

Developing CERs to Estimate Commercial IaaS Costs for Federal IT Systems

Cara Cuiule, PRICE® Systems LLC
Richard Mabe, PRICE® Systems LLC
Amanda Ferraro, PRICE® Systems LLC
Dan Harper, The MITRE Corporation

Abstract

Establishing federal budgets for cloud infrastructure costs prior to selecting a cloud provider requires vendor agnostic cost estimating methods. These methods need to reflect the correlation between rates for a variety of infrastructure instances across all viable cloud service providers. This paper describes research and validation leading to CERs and models based on over 28,000 virtual machine and storage instances. The predictive analytic approaches presented in this paper can provide valid and verifiable vendor agnostic estimates.

Table of Contents

1. Introduction	3
A. Purpose and Overview	3
B. Background	3
C. Virtual Machine (VM) Instances Definitions	4
D. Storage Definitions	5
E. Previous Work on Developing Cloud Pricing Models	7
F. Current Cost Estimating Issues	8
G. Applying Predictive Analytic Models as a Solution	8
2. Scope	9
A. Overview	9
B. Pricing Model Notional Format.....	10
3. Approach to Develop Predictive CERs/Models.....	10
A. Virtual Machine CER Development.....	11
B. Storage Pricing Models	19
4. Overall Model Validation and Comparison.....	22
A. Agency Model Comparison Process.....	23
B. Metrics	24
C. Cases	24
D. Overall Validation Conclusions	39
5. Challenges and Next Steps.....	40
A. Challenges	40
B. Next steps	41
6. Conclusion.....	41
7. Bibliography	42
8. Attachment A: Detailed Statistical Results for Each CER	44

1. Introduction

A. Purpose and Overview

Federal government agencies are using commercial Infrastructure as a Service (IaaS) to host their Information Technology (IT) systems in the cloud. When preparing to migrate their IT to a cloud host, these agencies need to establish cloud computing budgets as much as two years in advance of selecting a cloud vendor. This paper presents CERs and models for estimating IaaS costs to support budget development when a specific vendor is not known (considered in this paper as “vendor agnostic”). The paper focuses on predicting compute and storage costs for multiple IaaS instances, representing different pricing structures and services for six of the leading commercial cloud providers. This is a follow-on study to the research “Using Predictive Analytics and Open Source Data to Estimate IT and Cloud Related Costs for Government IT Systems”, which was presented at the 2019 ICEAA Workshop [1].

The paper first provides background on cloud computing, including research on cloud IaaS pricing for virtual machine computing services and for cloud data storage services. For each type of cloud service studied, this paper describes the data collection, normalization and analysis processes leading to predictive cost methodologies. Data was collected from the major cloud providers: Amazon AWS [2], Microsoft Azure [3], Alibaba [4], Google [5], IBM [6], and Oracle [7]. Six compute pricing CERs were developed along with five storage pricing models. Results from these proposed models are then verified using randomly selected datapoint values from the original dataset to validate the quality of the predictive methods. They are also compared to vendor-specific federal agency IaaS cost estimating models. The paper then presents conclusions on the usefulness of the models/CERs for developing multi-year budget estimates.

B. Background

The Federal Data Center Consolidation Initiative (FDCCI) enacted in 2010 directed federal agencies to audit, analyze and reduce the number of federal IT datacenters around the world [8]. It is now superseded by the Data Center Optimization Initiative, which also fulfills the requirements of the Federal Information Technology Acquisition Reform Act of 2014 to essentially overhaul federal IT [9]. Under the guidance of these directives, all agencies and departments of the federal government are transitioning much of their IT computing and data storage services to cloud infrastructure as a means to save money and improve operating capability. The cloud centers replace aging and expensive infrastructure at government data centers, allowing federal agencies to reduce the number of organic data centers and their associated operating, capital improvement and facility costs.

However, these agencies must now budget for and pay the recurring fees and expenses for the cloud services, in place of the labor, material and facility costs associated with their organic centers. Agencies must now develop and execute cloud budgets, based on defensible cost

estimates. These estimates must account for the varying rates and services provided by multiple commercial vendors, each offering a wide variety of infrastructure services.

The next section describes some of the basic definitions for cloud compute and storage services, the two services studied in this paper.

C. Virtual Machine (VM) Instances Definitions

The compute environment consists of the following infrastructure items that make up a virtual machine. [Figure 1](#) presents the relationship of these items in a compute and storage configuration. Separate storage as shown in this figure is discussed in the next section. Definitions are extracted from Wikipedia [10] and commercial vendor web sites [2] – [7].

Virtual Machine (Compute) An emulation of a computer system. Their implementations may involve special hardware or software. The components/hardware that make up a typical virtual machine are described below:

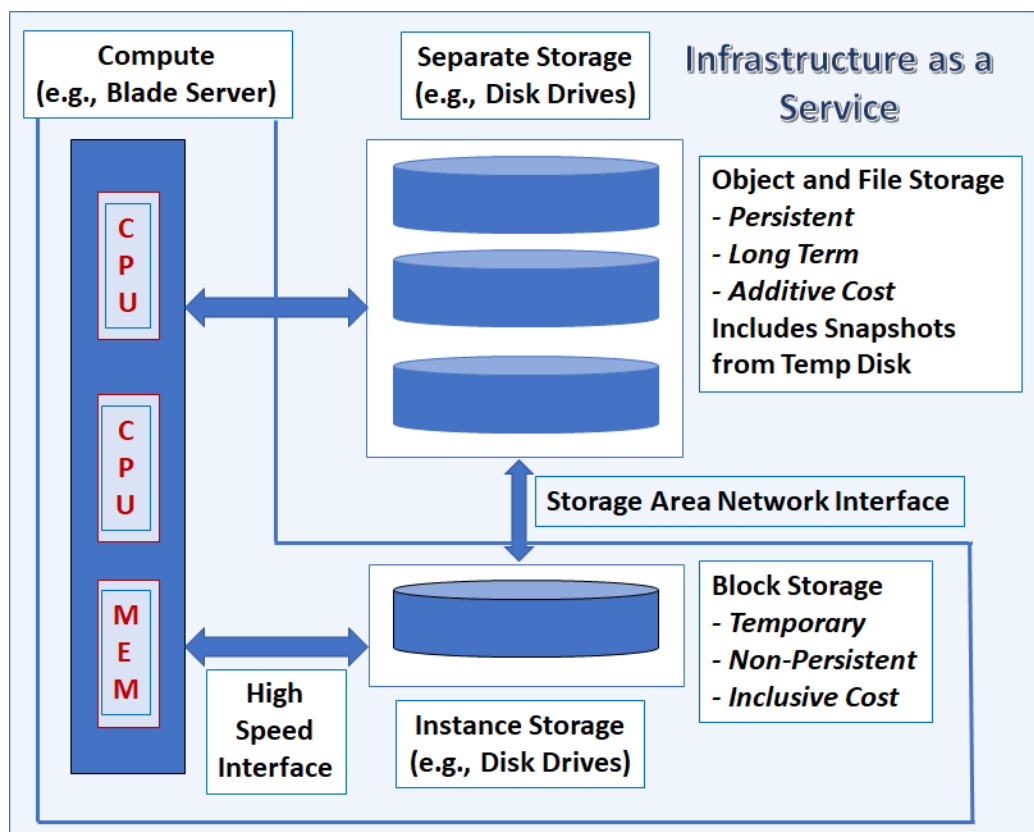


Figure 1. Compute and Storage Environments

CPU (Central Processing Unit) A circuit-based component that executes the most basic instructions of a computer program. It performs tasks such as basic arithmetic, logic, controlling, and input/output (I/O) operations.

MEM: Random Access Memory (RAM) This is computer memory and is typically used to store working data and code that is used by the CPU. It can read or write data items at the same rate regardless of the physical location of the data.

Cloud Network A computer network that is specifically for a cloud infrastructure. It enables the connections between cloud based or enabled applications, services and solutions. In the figure, this is represented by the interfaces (arrows) between the VM and storage.

Bandwidth The maximum rate of data that can be sent across the network.

Hourly Pricing (\$/Hour) The fee structure to recover user access to the compute resources can be paid the following ways depending on the provider: full upfront, no upfront, or partial upfront payments. Virtual machine hourly compute rates are generally based the number of CPUs and the amount of RAM memory (in Gigabytes) per virtual machine, and the overall number of virtual machines. There are multiple payment models, all of which won't be covered in this paper. The two specifically used for CER development are defined below:

On-Demand (Pay as you go) A business model used as the basis for computing CERs. Computing resources are made available to the user on an "as needed" basis, rather than all at once. This allows cloud hosting companies to provide access to computing resources as they become necessary.

Reserved Rate Discounts Reserved Instances are a billing discount applied to the use of on-demand instances for a contracted period of time that may give significant discounts. There are other types of discounts for virtual machines, but this study only applied one-year and three-year reserved instance discount rates to generate separate CERs from the hourly on-demand rates. Note that not every provider refers to these discounted rates by this term (for example, Google labels this type of pricing as "committed").

D. Storage Definitions

Research identified three types for storage capability incorporated with cloud computing: block, object and file. All of these types of storage can be purchased separately from virtual machines. Block storage capability is sometimes included in a virtual machine's hourly price and did influence CERs.

Block Storage Data that has a maximum ("block") size. Each block is assigned an arbitrary identifier but no metadata. It can be stored or retrieved a whole block at a time, which can boost reliability and efficiency. Block storage devices can be separate disk drives.

Note: As stated above, this research assumes that block storage is the type of storage used in conjunction with specific virtual machine instance rates. Block storage in a compute instance is not saved when the instance is closed, but a snapshot of the data state in the instance can be saved as file storage. Then it is retrieved as incremental backups, restore points, long-term storage, or as starting points for new Cloud Block Storage (CBS) volumes.

The other storage categories available for cloud users, object and file, are also available during a compute session, but the rates for the storage capacity are never part of the hourly compute instance and are always billed separately:

Object Storage Object storage differs from the other types of storage because it treats data as individual units (objects). Each object usually has accompanying metadata and unique identification. Objects can include pictures, music, videos or other media. This type of storage device can be controlled by an external operating system, but the data itself can only be accessed through APIs or web interfaces. It can be used for data backups.

File Storage File storage contains data in a hierarchical structure (files and folders). Files are accessed through shared systems, and do not require separate APIs. Like object storage, file storage is for unstructured data. It is also called file-level or file-based storage.

Below are additional definitions also associated with storage:

Hard Disk Drive (HDD) HDD uses magnetism to read and write digital data.

Solid State Drive (SSD) SSD is flash memory storage that uses integrated circuit assemblies to read and write digital data. It has much faster bootup speed compared to HDDs.

Snapshots This is a type of backup that preserves a Virtual Machine instance and its data at a specific point in time without any downtime. They can be a part of file or block type storage.

Network Definitions In addition to storing the data, network operations are typically used to manage and transmit file and object storage data:

Data Transfer Data can be downloaded from the cloud or moved to another location such as the internet or another data center.

Operational Requests Data operations within storage. Request operations include copying data, uploading an object, downloading an object, and returning certain pieces of information about that object.

Pricing Typically, storage is charged per GB of use (pay as you go) or as an upfront monthly charge for a set amount of GB. An exception to this includes operational requests, which are usually purchased as bundles of 1,000 or 10,000 requests.

E. Previous Work on Developing Cloud Pricing Models

While the cloud providers studied in this research have pricing calculators online that generate estimates based on their current prices, other models to predict cloud pricing have also been built by third parties. For example, Portella et. al. developed models for two types of AWS compute pricing (spot and on-demand). The model most relevant to this paper, the multi linear regression based on the amount of virtual CPUs, RAM, and ECUs (an AWS specific measure of processor power) was based on 2014 and 2016 data. However, they only validated their equation against six AWS memory-optimized type instances [11]. Smirnoff and Souiri of the National Reconnaissance Office Cost and Acquisition Assessment Group (NRO CAAG) have built models to forecast and calculate federal AWS compute, storage and database pricing based on historical trends [12]. Zhang et. al. developed software called “CloudRecommender” that would compare cloud service pricing based on requirements specified by the user [13].

The authors also obtained other models developed by a variety of federal agencies to support estimating the fees and service costs associated with commercial cloud virtual machines and storage environments were reviewed. To maintain the sensitive nature of these models, specific examples are not presented here. However, all of the models exhibited similar characteristics. For example, they were ad hoc tools developed and executed using Excel spread sheets and built-in numerical functions. These sheets included data tables and service fee rates provided by at least one of the following commercial cloud vendors: AWS, IBM, and Azure. The tables did not include all vendor provided IaaS options, only a selected set of options and associated rates. IaaS costs were then estimated using the rates as they applied to fixed instance configurations.

These models essentially offered results similar to online calculators for the vendor web sites (recurring fees by defined instance) but included additional estimating methods for transition and support costs. Some of the models also included approaches to compare rates between multiple vendors, to help determine the best-value approach for the defined requirements of the agency. They will be described further in [Section 4](#).

While no true “vendor agnostic” models were found during this initial research, it is important to note that the research to find already existing models was by no means exhaustive. The next section discusses the methodology used to develop a “vendor agnostic” model.

F. Current Cost Estimating Issues

The existing models and vendor web sites described above provided excellent data and methods to determine near term costs (current budget year, next budget year) for cloud services. They all assumed a specific commercial vendor would be used, and the rates were specific to that vendor and their advertised instance configurations. Users had to select a specific vendor and configuration to determine potential cloud costs and fees.

However, these models might not work for all situations given that the federal budget development and approval process is lengthy and complex. Costs supporting user requirements are estimated anywhere from 1 to 5 years in advance of when the funds will be executed. Additionally, cloud services and their rate structures are not standardized among providers. It is possible that agencies seeking to migrate IT services to the cloud will not have selected a cloud vendor prior to needing a defensible estimate of what the cloud services may cost.

Another problem with estimating cloud computing prices is their dynamic nature, thus making any model created irrelevant after a period of time. Pricing is market-dependent, and there is no predictable schedule for changes. Smirnoff and Souiri [12] studied AWS compute and storage pricing changes over 11 years. They found that the pricing changes were almost always price decreases that were 10 months or more apart (except for one price change that happened after a 2-month period). Every year, storage prices dropped about 12% and 6% percent for virtual machines. RedMonk also had a series of articles from 2018 [14] and 2019 [15] that included a dataset of 53 total virtual machines from a single region. Two providers (IBM and Oracle) decreased prices by an average of 23%. The other four providers did not change prices at all. Another online post, by Eric Lu, demonstrated with a small dataset that over several years Google storage and compute dropped about 5% in one region from 2016 -2018 [16].

Overall, there have not been thorough studies on pricing trends within all providers. However, it is sensible to expect general yearly decreases in prices as computing storage components becomes cheaper [12]. While average pricing changes are shown to be generally small, even a 5% change for an hourly cost can have a substantial effect over a longer period of time. Thus, it is important for the data behind any model created to be updated *at least* yearly.

These conditions create a complex and time-consuming scenario for cost estimating: evaluating and reporting on thousands of possible configurations from multiple vendors to determine a best-value 1 to 5 years into the future. There is a need to provide cost estimators with tools and methods that will deliver defensible estimates of this complexity in a timely and efficient manner.

G. Applying Predictive Analytic Models as a Solution

The complexity of vendors, instances and fees can be captured and evaluated using data tools to extract, transform/normalize and load thousands of possible instance rates from multiple

vendors into a searchable, standardized database. This data can then be filtered, scaled and correlated to create predictive cost models that support future budgets for cloud compute and storage requirements. Given the uncertainty of what specific vendor will be chosen, and what specific instances will be available, the best-fit approach for a model should include the following characteristics:

- Provide a non-vendor-specific rate, within an acceptable range of any vendor's rate
- Based on a solution that is similar to a vendor's online calculator or listing of offerings, so as to make the new model's design more familiar for a user:
 - Number of virtual machines (VMs)
 - Number of core processing units (CPUs) and amount of random-access memory (RAM) per VM instance
 - Amount of readable memory that can be rapidly accessed by the processors for computing (block storage)
 - Plus, the cost of additional storage not required to be tied directly to a compute instance (object and file storage)
 - Add the expected cost for recurring user access with the cloud instances and storage (networks)

This paper describes non-vendor-specific compute models (developed as CERs) and storage models (factors, benchmarks) that meet the need for complex cost analysis of commercial cloud rates and fees leading to defensible budgets.

2. Scope

A. Overview

The models described here best fit business IT systems such as finance, human resources, medical, and other sectors whose pricing will mostly resemble the commercial data available online. The largest percentage of the rates used in the analysis were for general application; however, a small portion of the rates represented costs for government-only commercial cloud centers. All rates were grouped together for the analysis.

The methods shown here are for commercial providers only. Government cloud providers, such as the Defense Information Systems Agency (DISA), and agency private cloud centers are not included. Data was extracted from web sites for six of the most common commercial cloud providers and includes over 28,000 world-wide data points for compute configurations as well as US data for storage configurations. Additional filtering, scaling and normalization to achieve CERs will be described later in the paper.

This research is part of a larger effort by the authors to capture and model the costs for a complete cloud transition and operations life cycle. However, this paper is limited just to a large subset of costs for IaaS, further described below.

B. Pricing Model Notional Format

The above descriptions provide a map to organize and present cost estimating methods from the commercial cloud vendor web sites. Five basic model formats provide a framework for presenting costs by compute instance, with and without block storage disks, and modified by additionally priced object and file storage. The table below summarizes each format.

Essentially, each instance price (or fee) will consist of:

1. A compute configuration (VM + Mem)
2. Plus, a combination of (or none of) the following storage configurations:
 - a. Included Block Disk
 - b. Additional Object and File Storage
 - c. Additional Snapshot file of a block configuration to store on an object disk

Pricing by instance is shown in [Table 1](#). Only the block storage rates are included in the compute rate. All object and file storage rates are added to the compute rate.

Table 1. IaaS Model Formats

Model	Infrastructure Components	Explanation
<u>Basic Infrastructure Models</u>		
Format 1	Number of CPUs + Memory (GBs)	Basic VM without Storage
Format 2	Added Storage	Basic Disk (Block, Object, File)
<u>Combined VM and Storage Instances</u>		
Format 3	CPUs + Mem + Temp Block Disk	VM + Included Storage; One Rate
Format 4	CPUs + Mem + Separate Object/File Storage	VM + Added Storage; Multiple Rates
Format 5	CPUs + Mem + Temp Block Disk + Snapshot Storage File	VM + Included Storage + Separate Snapshot; Multiple Rates

3. Approach to Develop Predictive CERs/Models

This section describes the process to evaluate commercial provider data and develop cost estimating methodologies for the compute and store environments. The objective of the analysis was to discover correlated relationships between “On-Demand Hourly” price and other database values that could be used to develop cost estimating relationships (CERs) to predict “On-Demand Hourly” price.

Each environment is discussed separately but includes the same level of detail on the data, analysis, and model building. Validation and comparison to other models are discussed in the next section.

A. Virtual Machine CER Development

This section discusses the compute environment, which may include block storage as part of a compute instance hourly rate.

1) Compute Environment Data

Data collection was a significant effort as thousands of datapoints and supplemental information was gathered from multiple sources and normalized to a format more digestible for statistical programs such as Excel, TrueFindings®, and ACEIT. Much of the virtual machine data was gathered through two third-party websites: Banzai Cloud [17] and Amazon EC2 [18]. It was then combined using a multi-step Extract-Transfer-Load (ETL) process.

Banzai Cloud provided key info for Linux virtual machines. Linux is open source so there would be no extra cost for this operating system, thus indicating a good baseline price to compare between different providers. Additional information included configuration name, number of CPUs, number of Graphical Processing Units (GPUs) if applicable, Memory, Network, and on-demand pricing. It also included “spot pricing” where instances are delegated based on supply and demand. This was deemed too dynamic to include with this analysis.




CATEGORY	MACHINE TYPE	CPUS	MEMORY	NETWORK	ON-DEMAND-PRICE (LINUX)
	VM.Standard2.1	1 vCPUs	15.00 GB	1 Gbps	0.06380 \$
	VM.Standard.E2.1	1 vCPUs	8.00 GB	0.7 Gbps	0.03000 \$
	VM.Standard.E2.2	2 vCPUs	16.00 GB	1.4 Gbps	0.06000 \$
<pre>{ "products": [{ "category": "Memory optimized", "type": "VM.Standard2.1", "onDemandPrice": 0.0638, "spotPrice": null, "cpusPerVm": 1, "memPerVm": 15, "gpusPerVm": 0, "ntwPerf": "1 Gbps", "ntwPerfCategory": "medium", "zones": ["sAkO:EU-FRANKFURT-1-AD-1", "sAkO:EU-FRANKFURT-1-AD-2", "sAkO:EU-FRANKFURT-1-AD-3"], "attributes": { "cpu": "1", "instanceTypeCategory": "Memory optimized", "memory": "15", "networkPerfCategory": "medium", "currentGen": false } }, { "category": "Memory optimized", "type": "VM.Standard.E2.1", "onDemandPrice": 0.03, "spotPrice": null, "cpusPerVm": 1, "memPerVm": 8, "gpusPerVm": 0, "ntwPerf": "0.7 Gbps", "ntwPerfCategory": "low", "zones": ["sAkO:EU-FRANKFURT-1-AD-"] }] }</pre>					

Figure 2. Comparison of Banzai Cloud website to JSON cURL code used for data collection

The Amazon EC2 website’s data only focused on that specific provider and was used to supplement what was already provided by the Banzai Cloud site. We collected “No Upfront” payment type reserved pricing, and local attached disk properties where applicable. While data collection was accomplished through web scraping, it is important to note that the website also allowed data to be downloaded from one location at a time to a “.csv” file type, a simple file format which can easily be opened in Excel. BanzaiCloud had data in a JavaScript Object Notation (JSON) file format that could be accessed through a cURL [19]. Data was extracted and then transformed into a more readable dataset using Rapidminer [20]. Amazon EC2 (compute) data was web scraped with a program called Octoparse [21]. Web scraping entails software

("bots") that build repeatable processes to more easily gather data. (For more information about the team's web scraping best practices, reference the author's previous ICEAA paper "'Big Data' Analytics in Operations Research" [22].) All datasets were finally combined using Excel.

However, there were some other helpful specifications about virtual machines that were missing from these two sources, such as local disk size (if included) and reserved instance pricing. We also collected much of our data through the vendors themselves: Amazon, Microsoft Azure, Alibaba, Google, IBM, and Oracle. Many of these websites had anti-web scraping policies listed in their Terms of Service pages, so the data from most of these websites were collected in a non-automated fashion.

Overall, 27,000 datapoints were collected for virtual machines. A large majority of them were for commercial instances, but Microsoft Azure did have about 1,700 Government instances available in two locations listed on their website [3].

The categories for the data are as follows:

- | | | |
|----------------|----------------------------|-----------------------|
| • Name | • GPU Type | • Network Performance |
| • Type | • Local Disk (If included) | • Provider |
| • Category | • Hourly Prices | • Continent |
| • CPUs | ○ On-Demand Price | • Region |
| • RAM (Memory) | ○ 1 Year Reserved Price | • Location (City) |
| • GPUs | ○ 3 Year Reserved Price | • Date |

2) Data Normalization

All of the data extracted for this analysis were then transformed (normalized) to a common numerical format and tabular format (Excel worksheets) for analysis. This transformation included the following:

- a. Convert all "text" values that describe a numerical value to numbers. [Table 2](#) provides some examples of local disk size values taken from web sites, that are stated as "text", and their corresponding numerical value:

Table 2. Converting Text to Numerical Values - Examples

Example Instance	Local Disk Size (Text)	Local Disk Size (Number of Gigabytes)
1	1 x 1250 NVMe SSD	1,250
2	1 x 7500 NVMe SSD	7,500
3	1 x 120 SSD	120
4	12 x 2000 HDD	24,000
5	2 x 900 NVMe SSD	1,800

Several of the Network Performance values were also expressed qualitatively (Very Low, Low, Low to Moderate, Moderate, High and Very High). The ranges of all instances were evaluated, both qualitative and numerical, and resulted in the following rule for converting qualitative values to numerical values, based on the distribution of the entire dataset.

Table 3. Converting Text to Numerical Values - Examples

Example Instance	Network Performance Throughput (Text)	Network Performance Throughput (Number of Gigabytes)
1	Very Low	1
2	Low	10
3	Low to Moderate	20
4	Moderate	35
5	High	50
6	Very High	100

- b. Convert numerical size measured by the commercial vendors as “Gibibytes” (GiB) to “Gigabytes” (GB). A GiB is a computing measure, used in binary computing systems for measuring data; where:

$$1 \text{ GiB} = 2^{30} \text{ bytes} / 1,000,000,000 = 1.074 \text{ GB}$$

- c. Adjusting for skew. The “On-Demand Hourly” price values in the data set exhibited a strong right skew when plotted. To mitigate the impact of the skew on this analysis, all numerical data was transformed using the natural log function (LN) in Excel:

$$\text{Log Transform Value} = \text{LN}(\text{Database Value})$$

The following figures show the non-transformed on-demand hourly rates with a high skew, and the transformed database. The transformed data, distributed around the

lognormal mean, were used to identify correlated data pairs leading to subsequent lognormal CERs:

$$\text{LN}(Y) = B + \text{LN}(X) * m$$

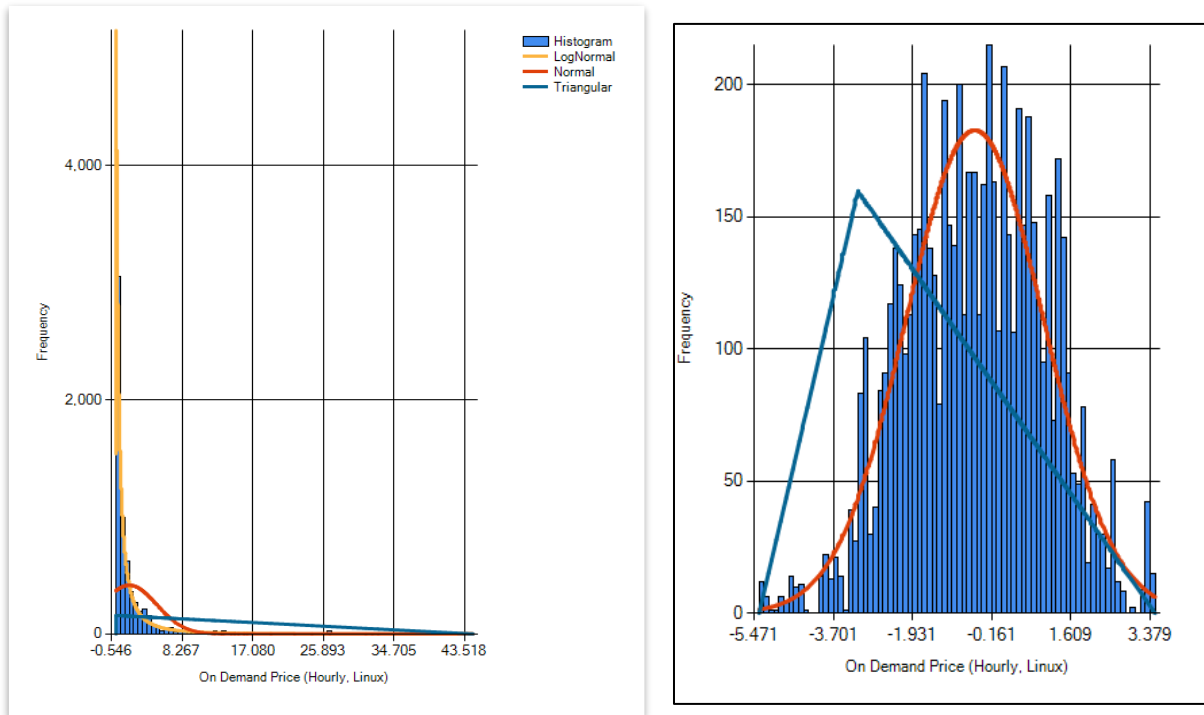


Figure 3. As-Is Database (on the left) and the Transformed database (on the right)

3) Filtering of Normalized Data

Prior to evaluating the normalized data for correlated pairs, additional steps were taken to filter the data and focus the analysis only on specific data characteristics. Starting with 26,645 valid instances, the following actions resulted in a dataset of 6,391 instances to use for determining CERs:

- Removed 17,089 instance rows with a blank value for the key elements to be tested for correlation: number of CPUs, compute memory (RAM), Local Disk (e.g, block storage) and on-demand rate. Blanks were interpreted as an error, and not as a value of “zero”. However, if the value in a cell was intentionally equal to zero, we counted that as a value and not a blank.
- Removed rows with blanks for key values also resulted in a reset of the upper and lower boundaries for the dataset. The tables below provide snapshots of the final dataset configuration and the statistical profile for On-Demand hourly rates:

- c) Included only instance rates and associated data for US commercial cloud locations. Later analysis may focus on specific non-US regions, such as Europe, South America, Canada, Mexico, Australia and Asia, to develop localized CERs. The final count was 6,391.

Table 4. Dataset Configuration for US Locations (Count = 6,391)

Data Element	Lower Boundary	Upper Boundary
Number of CPUs	1	128
Compute Memory (RAM) (GB)	0.5 GB	4,080 GB
On-Demand Hourly Rate	\$0.0047 / Hour	\$32.76 / Hour
Local Disk	0 GB	60,000 GB

Table 5. Statistical profile for On-Demand Hourly Rates, US Locations (Count = 6,391)

Statistic	Value	LogNormal
Min	0.005	
Max	32.760	
25%	0.191	
75%	1.872	
Mean	1.711	1.938
Median	0.610	0.589
Mode	0.043	0.054
Standard Deviation	3.278	6.073
Coefficient of Variation	1.916	3.133

- d) The database was filtered in different ways to develop the different CERs:
- For the non-discounted base rate with local disk space, there were only 2,884 total datapoints included from 2 commercial vendors (Microsoft and Amazon).
 - For the separate one and three year discounted reserved rate CER development, we removed additional instances with blank values for the discount:
 - This left only 2,215 instances to evaluate for discount rates for the CPU + MEM CERs, which only included 3 commercial vendors (Microsoft, Amazon, and Google).
 - For the discount CERs with local disk space, we filtered the database down to 1,161 datapoints from 2 commercial vendors (Microsoft and Amazon).

4) Developing CERs for Compute Environment Pricing

The following actions describe the process to evaluate the normalized data and develop CERs to predict On-Demand Hourly pricing for compute instances, first for the regular rate, then for discount rates.

i. Data Correlation

Figure 4 presents the correlated data pairs from the “as is” (original) database and the Log Transform database. Data pairs were checked for correlation using Pearson’s Correlating Coefficient (R). A value of 0.70 or higher represented significant correlation between data elements.

On-Demand Hourly rates were most closely correlated with the amount of compute memory required (RAM). Lesser correlation (> 0.50 , < 0.70) occurred between the on-demand rates and the number of CPUs supported in an instance. There was no significant correlation between instance rates and the network connecting virtual machines and data in a data center.

For the original database (pre-Log transform), there is no significant correlation between the potential independent variables that would lead to multi-collinearity in a CER. However, in the Log transform database, there is significant correlation between the number of CPUs and the RAM in an instance.

As-Is Database

Dependent =	F(Independent)	R
On Demand Price (Hourly, Linux)	Memory (GB)	0.855
On Demand Price (Hourly, Linux)	CPUs	0.585
On Demand Price (Hourly, Linux)	NW Performance-Numerical (GB)	0.125
Memory (GB)	CPUs	0.522
Memory (GB)	NW Performance-Numerical (GB)	0.029
CPUs	NW Performance-Numerical (GB)	0.218

After Log Transform

Dependent =	F(Independent)	R
On Demand Price (Hourly, Linux)	Memory (GB)	0.959
On Demand Price (Hourly, Linux)	CPUs	0.887
On Demand Price (Hourly, Linux)	NW Performance-Numerical (GB)	0.091
Memory (GB)	CPUs	0.861
Memory (GB)	NW Performance-Numerical (GB)	0.075
CPUs	NW Performance-Numerical (GB)	0.214

Figure 4. Correlation Within the Database

ii. CER Configuration and Selection

Potential CERs were evaluated for the As Is (Unit Space) and Log Transform (Log Space) database, using single and multiple regression approaches, first for a basic rate, then for

discount rates. [Table 6](#) below presents the model formats evaluated through regression modeling to develop a CER. Results for each CER were then evaluated by comparing the predicted values from the CER to the actual values in the dataset, using the following metrics:

- Coefficient of Correlation (R^2) – Values > 0.70 were assumed to be a good fit
- F-test for the Model – a P-value less than alpha for the one tailed test (0.10) implied the CER model was a better predictor of on-demand rates than the mean of the “On-Demand Hourly Rate” distribution for all instances.
- t-Test for the model variables and model intercept - a P-value less than alpha for the two-tailed test (0.05) implied the variables used in the CER model were valid predictors of on-demand rates, providing a better fit than using just the mean of the on-demand hourly rate distribution for all instances.

Table 6. Model Formats Evaluated Through Regression

Model	Infrastructure Components	Explanation
<u>Basic Infrastructure Models</u>		
Model 1	Number of CPUs + Memory (GBs)	Basic VM without Storage
Model 2	Number of CPUs + Memory (GBs)	Basic VM – 1 Year Reserve Discount
Model 3	Number of CPUs + Memory (GBs)	Basic VM – 3 Year Reserve Discount
<u>Combined VM + Storage Instances*</u>		
Model 4	CPUs + Mem + Temp Storage + Snapshots	Basic VM + Storage
Model 5	CPUs + Mem + Separate Object/File Storage	Basic VM + Storage (1 Year Discount)
Model 6	CPUs + Mem + Separate Object/File Storage	Basic VM + Storage (3 Year Discount)

**Separate (Object and File) Storage Models evaluated in paragraph 3B. This table is for block storage only.*

The results (“best-fit” models) are summarized below. Details for each metric by CER model are presented in [Attachment A](#). All models are compared in Unit Space for consistency.

Lognormal models provided the best-fit for estimating on-demand hourly rates that captured the skew of the database. Multi-regression models, combining the number of CPUs, the amount of RAM and the amount of local disk storage (block storage) provided the best overall fit by instance for the basic rate and discounted rates. However, the equations needed to be converted back to unit space for rates to be in regular currency and not the lognormal currency by raising both sides to the exponential function. Note that the R^2 is for the unit space equation. Additionally, the “RAM” (Memory) and “Local Disk” variables are in GB:

Table 7. Summary of CERs for On-Demand Hourly Rate

Eq #	CER in Unit Space (Hourly Cost per VM)	R ²
VMs Without Local Disk		
1	Rate = exp(0.662 * LN[RAM] + 0.337 * LN[vCPUs] – 3.680)	0.71
2	1 Yr Disc. Rate = exp(0.679 * LN[RAM] + 0.356 * LN[vCPUs] – 4.367)	0.67
3	3 Yr Disc. Rate = exp(0.663 * LN[RAM] + 0.370 * LN[vCPUs] – 4.731)	0.63
VMs With Local Disk		
4	Rate = exp(0.602 * LN[RAM] + 0.224 * LN[vCPUs] + 0.146 * LN[Local Disk] – 4.029)	0.83
5	1 Yr Disc. Rate = exp(0.630 * LN[RAM] + 0.278 * LN[vCPUs] + 0.080 * LN[Local Disk] – 4.433)	0.79
6	3 Yr Disc. Rate = exp(0.588 * LN[RAM] + 0.270 * LN[vCPUs] + 0.120 * LN[Local Disk] – 4.891)	0.69

Validation and comparison of the equations in [Table 7](#) to data are discussed in [Section 4](#). The results for the discount rates represented only two of the six commercial providers. Given the limited application and the R² values <0.70, factors to use instead of the CERs were also developed for the discount rates as shown below. Unlike the CER models above, these rules of thumb were not tested and are instead recommended as an alternate approach:

Table 8. Discount Rate Factors: Cost = Basic Result x (1 - Discount Factor)

Overall Discount	Actual Rate (All Data)	Factor for Compute Only (No Local Disk)	Factor for Compute + Local Disk
1 Year Reserve	35.2%	40.4%	39.9%
3 Year Reserve	59.6%	60.1%	55.5%

The next paragraph reviews the steps taken to collect, normalize and analyze data for added storage rates. These were used to develop predictive methods for estimating added storage costs that can be added to compute instance rates to predict the total rate for required compute plus storage configurations.

B. Storage Pricing Models

1) Data Collection and Sources

The data in this section refers to storage instances that are purchased separately from a virtual machine instance. About 1,600 datapoints were collected for storage from all regions in the United States. Standard storage prices were selected and were labeled by type: Object, File, and Disk, and Snapshot. This data was collected (by hand) directly from the six provider websites.

The categories for the datapoints are as follows:

- Name
- Storage Type (Disk, Snapshot, Object, File)
- Disk Type
- Storage Size (GB)
- Monthly Price/TB
- Monthly Price/GB
- Provider
- Continent
- Region
- Location (City)
- Date

Name	Storage Type	Disk Type	Storage Size (GB)	On Demand Monthly Price (\$/TB/Month)	On Demand Monthly Price (\$/GB/Month)	Provider	Continent	Region
File Storage - Google - us-east1	File	-	-	204.8	0.20	Google	North America	us-east1
File Storage - Azure - Central US	File	-	-	297.984	0.291	Azure	North America	Central US
Object Storage - Amazon - US West - Oregon	Object	-	-	36.152	0.03530	Amazon	North America	US West
Object Storage - Azure - East US - Virginia	Object	-	-	24.8992	0.02432	Azure	North America	East US
B1.1x2x25	Disk	SSD	150	78.0288	0.0762	IBM	North America	USA
Cloud Disk	Disk	SSD	100	18.944	0.0185	Alibaba	North America	US East 1
S15	Disk	HDD	274.87744	42.1888	0.0412	Azure	North America	West US
Ultra Cloud Disk	Disk	HDD	100	20.48	0.02	Alibaba	North America	US East 1
Snapshot per region	Snapshot	-	-	51.2	0.05	Azure	North America	Central US
EBS Snapshots to Amazon S3	Snapshot	-	-	56.32	0.055	Amazon	North America	US West

Figure 5. Snapshot of Storage Pricing Data

2) Normalization

Similar to the virtual machine data, storage data was also normalized to a common format for implementation and analysis. This transformation included the following actions:

a. Converted numerical size measured by the commercial vendors as “Gibibytes” (GiB) to “Gigabytes” (GB). Also, “Gigabytes” (GB) were converted to “Terabytes” (TB) when necessary.

$$1 \text{ GiB} = 2^{30} \text{ bytes} / 1,000,000,000 = 1.074 \text{ GB}$$

$$1 \text{ TB} = 1,024 \text{ GB}$$

b. Normalized pricing structures. For File Storage, some providers gave a price for pay as you go, while others gave a price for a set amount of GB being stored. These prices were normalized to the amount of dollars paid per TB.

However, Alibaba’s structure was different than the rest. Instead of a monthly price per unit, their structure consists two payment methods. Each payment method was averaged across all

regions the provider offered. Then, the two payment methods were averaged together for an overall average for that provider.

For object storage, data was normalized into tiers based on all six providers. Operational request data was also normalized to a dollar rate per 10,000 requests. Below is the comparison of object storage pricing that clearly lays out the pricing tiers. Note that this format comes from an Enterprise Storage article [23]. The prices are from April 2018:

	Amazon S3	Microsoft Azure Blob Storage	Google Cloud Storage	IBM Cloud Object Storage	Oracle Cloud Object Storage
Free tier	5GB for 1 year	5GB for 1 year	5GB forever	25GB forever	5TB for 30 days
Price per GB	<ul style="list-style-type: none"> • 1st 50TB \$0.023/GB • Next 450 TB \$0.022/GB • Over 500 TB \$0.021/GB 	<ul style="list-style-type: none"> • 1st 50TB \$0.0208/GB • Next 450 TB \$0.020/GB • Over 500 TB \$0.0192/GB 	<ul style="list-style-type: none"> • \$0.02/GB 	<ul style="list-style-type: none"> • 0-499.99TB \$0.022/GB • 500+TB \$0.02/GB 	\$0.0255
Data transfer out	<ul style="list-style-type: none"> • First 1GB free • Next 9.999TB \$0.09/GB • Next 40TB \$0.085/GB • Next 100TB \$0.07/GB • Over 150TB \$0.05/GB 	Free for hot data	<ul style="list-style-type: none"> • 0-1TB \$0.12/GB • 1-10TB \$0.11/GB • 10+TB \$0.08/GB 	<ul style="list-style-type: none"> • 0-50TB \$0.09/GB • Next 100TB \$0.07/GB • Next 350GB \$0.05 • 500+TB price available on request 	Free

Figure 6. Object Storage and Transfer fees from Enterprise Storage Website

Normalization occurred so that averages could be taken across providers. Any discounts were small, so they were ignored. Some providers offered more granular pricing, while others had tiers with larger ranges. If the tiers didn't match, the smallest ranges were picked as the "normalized" range, and the providers with larger ranges had that range broken up and the price within that range duplicated for the smaller one. For example, in the chart above for data transfer, Amazon S3 would be considered to have a rate of \$0.09/GB for 0-10 TB. However, Google has tiers of 0-1 TB and 1-10 TB, so Amazon rate of \$0.09/GB would be split between tiers of 0-1 TB and 1-10 TB. This methodology was implemented in [Table 9](#) below.

3) Model Building

There was a small number of datapoints for each type of storage, and no strong correlation existed between price and any variables in the dataset. Thus, storage was modeled similarly to the way it is already calculated by providers: the amount of memory multiplied by the price per that memory unit. There was generally a negative correlation between the amount of data stored and the price per unit; as the amount of storage increases, the price per unit decreases, as shown above in [Figure 6](#). Many of the terms used below were previously defined in [Section 1.D](#).

a) Object Storage

Pricing structures for storage were not uniform among all providers. For certain providers, storage and transfer fees were reduced based on the number of terabytes of data.

The object storage model was developed to account for these different storage models. Discounts were ignored since they were negligible, especially considering the scale of terabytes of data being stored or transferred. Furthermore, the Enterprise Storage article [23] cited networking (transfer and operational request) prices as another driver of cost for storage. Operational requests perform actions within Cloud storage. “WRITE” requests are used to send data to a server to create or update a resource while “READ” requests are used to request data from a specified resource [24].

[Table 9](#) shows the detailed version of our object storage model as of October 2019. All prices are given as dollar per terabyte rates. For each provider, all US region prices were averaged. Finally, all provider prices were averaged for the rates used in the model:

Table 9. Object Storage Pricing

	Amazon	Microsoft	Google	IBM	Oracle	Alibaba	
Data Storage Prices (Per TB)							Avg Rate (Vendor Agnostic)
First 50 TB	24.32	19.7632	21.7088	22.528	26.112	19.712	22.36
Next 450	23.296	19.008	21.7088	22.528	26.112	19.712	22.06
500+ TB	22.272	18.2528	21.7088	20.48	26.112	19.712	21.42
Data Transfer Out Prices (Per TB)							
0-1 TB	92.160		122.88	92.16		77.824	96.26
1-10 TB	92.160		112.64	92.16		77.824	93.70
10-50 TB	87.040		81.92	92.16		70.656	82.94
50-150 TB	71.680		81.92	71.68		61.44	71.68
150+ TB	51.200		81.92	51.2		44.032	57.09
Operational Request Prices							
WRITE Requests (per 10,000) or Class A	0.051	0.051	0.05	0.05	0.003	0.016	0.0368
READ Requests (per 10,000) or Class B	0.004	0.004	0.004	0.004	0.003	0.001	0.0034

iii. File Storage

The File Storage model was developed based on the amount of data being stored and the monthly rate. Averages for each provider were taken from all the US region prices. As

previously mentioned, the monthly rate prices were also converted to terabytes. The providers had an average rate of \$255.32/TB:

Table 10. File Storage Pricing

	Amazon	Microsoft	Google	IBM	Oracle	Alibaba	Avg Rate
\$/TB	\$ 314.88	\$ 284.64	\$ 217.09	\$ 253.44	\$ 307.20	\$ 154.68	\$ 255.32

Networking prices (data transfer out and operational requests included in [Table 9](#)) might also be a part of file storage, but after looking at all provider websites it was determined that half of the providers don't have networking prices associated with file storage. Among the three providers that do include file networking, the information on online documentation and calculators is sometimes contradictory. This will have to be studied further.

iv. Additional Block Storage

Block Storage consists of two disk types, Hard Disk Drives (HDD) and Solid-State Drives (SSD). Because SSD disk drives are generally known to be more expensive than their HDD counterparts, those prices were kept separate for the additional block storage model. For the model, an average monthly price was calculated for each disk type across all providers:

Table 11. Additional Block (Disk) Storage Pricing

Disk Type	Avg Rate (\$/GB)
SSD	0.1171
HDD	0.0900

v. Snapshot Storage

Snapshot storage is usually associated with block storage but can be included with file storage. The snapshot data collected for the model is specifically used for block (disk) storage. More research will need to be done to determine the difference in price between snapshots for disk storage and for file storage. An average monthly price was calculated from all the data and is the value used in the model. Shown below is a small section of the data:

Table 12. Snapshot Storage Pricing

	Avg Rate (\$/GB)
Snapshot	0.0375

4. Overall Model Validation and Comparison

The compute environment models and storage models represent statistically valid approaches to estimate non-vendor specific rates for commercially provided IaaS support. The final step in the analysis to determine defensible approaches for cost estimating includes validation of the models using random datapoints from the model and comparison of the models to three

federal agency calculators. This paragraph provides an overview of these processes as applied to data from three federal government agencies.

The names of the agencies and the specific data used for comparison are purposely withheld. The intent is not to validate the *known costs* for any specific agency requirement. Rather, the intent is to use the approaches developed in this study to estimate known solutions and compare the CER results to the known solution results.

A. Agency Model Comparison Process

The figure below provides a visual overview of the process used to compare the models to federal data. The federal data included spreadsheet models developed by the agencies to estimate IaaS costs for their specific commercial requirements, and generally for a specific commercial vendor. We used the specific vendor requirements (VMs, CPUs, RAM, Block Storage, Added Storage) as inputs to the non-vendor specific CERs and models to estimate a solution. Each agency had also used their spreadsheet model to estimate a solution for their requirements. The CER estimated solution was then compared to the agency generated solution to determine how well the CERs were able to estimate a different model.

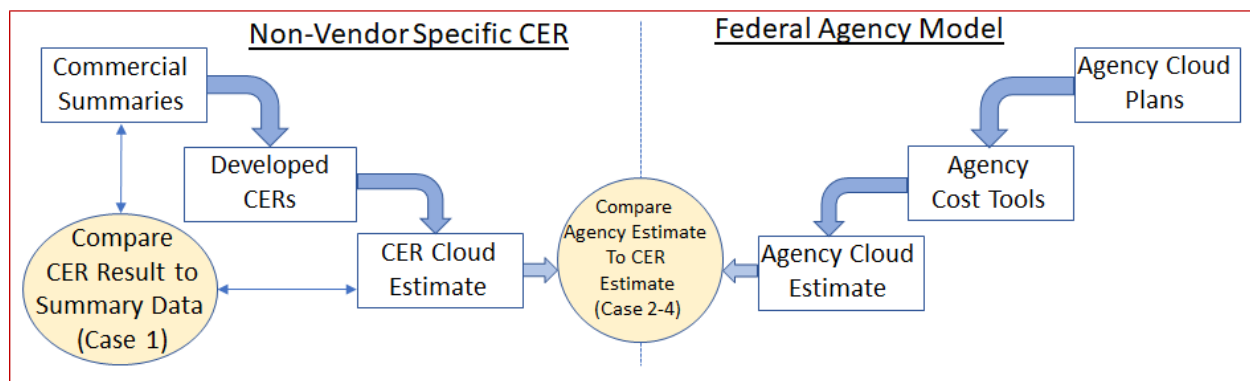


Figure 7. Agency Comparison Process

The following assumptions apply to this comparison.

1. Four specific cases were evaluated. The first was a validation of the dataset while the next three cases are federal agency calculator comparisons. Each case is discussed in Section C:
 - a. Compare the CER results to random instances from the data set (all CERs)
 - b. Compare the CER results to the results for Agency 1 (VM + MEM + Object Storage)
 - c. Compare the CER results to the results for Agency 2 (VM + MEM + Block Storage)
 - d. Compare the CER results to the results for Agency 3 (VM + MEM + Block Storage)

2. The agency estimated values were assumed to be a correct solution for the agency requirements. The comparison process was being used to evaluate the defensibility of the CERs and not to validate the “correctness” of the vendor solution. The authors do not know if the agency solution was actually implemented or funded.
3. If the CER and model generated solutions were not significantly different than an agency estimate, then we concluded the CER result was “as valid” as the agency result. In other words: **if the agency had used the non-vendor specific CER, they would have estimated a budget requirement of the same magnitude as estimated by their own agency model for a specific commercial vendor.**

B. Metrics

The following measures were used to determine if the difference was significant enough to declare the CER as not being a valid method to estimate on-demand hourly rates:

1. Average Absolute Percent Error, calculated as the absolute value of the Actual data point values minus the Estimated values using the CERs. A mean difference across all data points of less than one standard deviation (34%) for the complete data set was considered as not significant.
2. Pearson’s Correlation Coefficient (R), measured the linear correlation between the Actual data points and their corresponding estimated values using the CERs. A value of 0.70 or greater was interpreted as the two data sets showed significant correlation, meaning the difference between them was insignificant.
3. Coefficient of Determination (R^2), measured the amount of the variance in the estimated on-demand hourly rates (dependent variable) that is predictable from the independent variables (CPUs, VMs, MEM, Storage), as measured by comparing the actual on-demand hourly rate values to the estimated CER results. This is a measure of the quality of the CER. Values greater than 0.70 were interpreted as the variance between the two data sets (actuals, estimates) was insignificant, and that the independent variables were useful for predicting on-demand hourly rates (the CER is better than the database mean).

C. Cases

The following paragraphs explain the process to evaluate each case and determine if the CER estimate was not significantly different from prior agency estimated results. Each case below outlines the initial data in the models for Compute, Storage, and Networking prices. Then the process of how the case’s data was matched to the CERs/models developed in Section 3 was outlined (“Research Model Calculations”) and the two are compared (“Comparison”). Finally, results and conclusions are made. All cases tested virtual machine CERs, but not all included storage models. No model included file storage or snapshots.

1) Case 1: Compare CERs to Database

i. Compute Data

Out of the filtered database of compute datapoints, 100 instances were tested. These instances were randomly chosen from a subset of the data that had all three types of hourly pricing (on-demand, one year reserved and three year reserved). They were also selected so that 50 datapoints had disk storage attached and 50 did not. Due to a lack of diverse data for reserved instance pricing, this meant that only three providers were represented after this selection criteria (Amazon, Azure, and Google). However, these three vendors represented nearly 75% of the total dataset for the CER equations.

ii. Research Model Calculations

All six CERs from [Table 7](#) were used. The CER tested was based on the compute configuration. For example, to estimate an on-demand price for a VM instance without a local disk attached, Equation 1 was used. If it had a local disk attached, equation 4 was used instead. The appropriate equations (2,3,5 and 6) were used for discounted rates.

iii. Comparison

The table below presents the results of the CERs as compared to the database actual values:

Table 13. Results of using CERs on 100 random datapoints in the dataset

Metric	CPU + MEM Rate (CER)	CPU + MEM + Storage (CER)
On-Demand Rate		
Avg Abs % Error	37%	20%
R	0.67	0.97
R²	0.45	0.94
1 Year Discount Rate		
Avg Abs % Error	30%	23%
R	0.65	0.96
R²	0.42	0.92
3 Year Discount Rate		
Avg Abs % Error	27%	22%
R	0.73	0.94
R²	0.53	0.88

iv. Conclusions

The results for the **CPU + MEM + Storage** CERs (Equations 4-6) are good. The low average absolute error values (<25%) and high R and R² values (>0.70) reflect an insignificant difference between the Actual Database values and the Estimated CER values.

For the **CPU + MEM** CERs (Equations 1-3), the 3 Year Discount Rate fit is better than the Base Year and 1 Year Discount Rates. The estimated on-demand rate does show a significant difference for the Base and 1 Year Discount rates are not necessarily defensible CERs. This is especially true when considering that this is an hourly price, and the magnitude of the error will only grow as it is used to estimate prices over longer periods of time.

This was the first step in validation. Comparing the CERs to agency estimates will be used to verify these results, or to show any exceptions for the agency data.

2) Case 2: Agency 1

i. Agency Data

a) Compute Environment

Agency 1 developed an estimating tool that included a range of potential compute and storage configurations and associated fixed rates for each range based on AWS pricing. The ranges for virtual machine CPUs and associated RAM are shown. The rate per range is a fixed average that applies to any configuration of CPU and RAM within the ranges shown.

Table 14. Agency 1 Ranges and Suggested Rates for VMs

vCPU Range	RAM Range (assumed to be in GB)	Fixed Avg Monthly On-Demand Rate
16 to 32	128 to 256	\$2,037
8 to 32	64 to 128	\$1,197
8 to 32	32 to 64	\$956
4 to 16	16 to 32	\$402
4 to 16	8 to 16	\$314
4 to 8	4 to 8	\$241
2 to 4	2 to 4	\$183

b) Storage Rates

The Agency 1 model used fix monthly rates per GB range as shown below. Two types of additive storage were considered. There was a high-performance SSD supporting temporary block storage where the data drive is closely linked to the VM instance for high speed access, but the storage rate is not included with the compute instance rate. The standard Hard Disk Drive provides object storage in support of a user while maintaining the structure of the data.

Table 15. HDD and SSD Capacity and Rates Offered in the Agency 1 Model

Type	Storage (GB)	Fixed Average Monthly Storage Rate
Standard HDD	512	\$26
Standard HDD	1,920	\$94
Standard HDD	4,096	\$201
High Performance SSD	512	\$67
High Performance SSD	1,920	\$236
High Performance SSD	4,096	\$494

c) Network Rates

For network data transfer, two types of network activities were included: transfer out data for storage and operational requests for data (WRITE/READ SQL actions). The table below shows rates used by the agency for comparative analysis to determine if a cloud solution was economically viable (return on investment analysis).

Table 16. Network Data Transfer Rates for Agency 1

\$/GB/Month	Action	GB/Month	Hours	Total Cost
\$ 0.0950	Transfer Data (Out)	1,490	N/A	\$ 141.55
\$ 0.0300	Ops Data (WRITE/READ)	62,083	N/A	\$ 1,862.50

ii. Research Model Calculations

a) Compute CER

For every evaluated VM instance, there were three sets of values evaluated from each instance's vCPU and RAM range (low, mid, high) to evaluate with the VM hourly cost CER for CPU + MEM (Equation 1 in [Table 7](#)). The average monthly rate was calculated from the hourly rates assuming 720 hours per operating month (24 hrs x 30 days). The results for each option were then used to create a CER average monthly rate for each CPU/Memory range to compare to the Agency 1 model average monthly rate. For example, the first virtual machine offered in [Table 14](#) (16 -32 vCPUs, 128 – 256 GB RAM) would be estimated in the following way:

Table 17. Example of Virtual Machine Calculation Using CER Model

Position in Range	vCPU Qty	Memory (RAM)	\$/VM/Hr (Table 7 , Eq. 1)	Monthly Cost (\$)	Avg Cost Per Month (\$)
High	32	256	3.1863	2,294.14	1,721
Midpoint	24	192	2.3904	1,721.10	
Low	16	128	1.5943	1,147.86	

This process was repeated for all 7 virtual machine instances. The average was generally found to be within the same full dollar amount the “Midpoint” row.

b) Storage Rates

Agency 1 rates for storage had a fixed amount (512 GB, 1920GB, 4096 GB) and had both HDD and SSD offerings for each size. These were compared to rates developed for additional disk storage. The Agency 1 rates were derived from AWS vendor rates. The table below shows the database rates from [Table 11](#), representing the average for all vendors for the Agency 1 storage requirement:

Table 18. Agency 1 Disk Drive Pricing - Model Results

Type	Storage (GB)	Full Database Average Monthly Storage Rate
Standard HDD	512	\$46
Standard HDD	1,920	\$173
Standard HDD	4,096	\$369
High Performance SSD	512	\$60
High Performance SSD	1,920	\$225
High Performance SSD	4,096	\$480

c) Network Rates

The Agency 1 network rates per TB/month were compared to the average rates from the commercial vendor database ([Table 9](#)). Agency 1 was transferring 1,490 GB of data per month, which is equal to 1.455 TB.

Table 19. Agency 1 Transfer Pricing - Model Results

Network Model	Average/TB		
Transfer Out Prices (Per TB)	Average	Amt of TB at Price Range	Calculated Price
0-1 TB	\$ 96.26	1	\$96.26
1-10 TB	\$ 93.70	0.455	\$42.63
Total for 1,490 GB		1.455	\$138.89

The operational request prices are given per 10,000 requests and it is unclear how exactly these would map to the GB that the agency calculates:

Table 20. Agency 1 Request Pricing - Model Results

Operational Request Prices	Average (per 10,000 Requests)
WRITE Requests (per 10,000) or Class A	\$ 0.0368
READ Requests (per 10,000) or Class B	\$ 0.0034

iii. *Results Comparison*

a) *Compute CERs*

Here are the Agency 1 compute environment rates compared to the CER estimated rates for the compute environment along with averages across all ranges, and the associated quality measures comparing the Agency 1 model and the CER estimated results. The CER always underestimated the actual costs given by the agency.

Table 21. Agency 1 Compute - Model Results Comparison

vCPU Range	RAM Range per CPU (GB)	Agency 1 Avg Monthly On-Demand Rate	CER Estimated Avg Monthly On-Demand Rate	% Absolute Error
32 to 38	128 to 256	\$ 2,037	\$ 1,721	15%
10 to 32	64 to 128	\$ 1,197	\$ 1,016	15%
8 to 32	32 to 64	\$ 956	\$ 642	33%
4 to 16	16 to 32	\$ 402	\$ 321	20%
2 to 14	8 to 16	\$ 314	\$ 203	35%
4 to 8	4 to 8	\$ 241	\$ 109	55%
4 to 4	2 to 4	\$ 183	\$ 54	70%

Overall Avg CER Cost per Month	Overall Avg Agency Cost per Month	Delta Monthly Averages	% Average Absolute Error	R²
\$ 581	\$ 761	\$ 180	35%	0.99

b) *Storage Rates*

The table below presents a comparison of the Agency 1 object storage rates and the estimated rates from the vendor database for the complete database of 6 vendors.

Table 22. Agency 1 Disk Drives - Model Results Comparison

Type	Storage (GB)	Agency Monthly Rate	Database Avg Monthly Rate	Absolute % Error Database Rates
Std. HDD	512	\$26	\$46	79%
Std. HDD	1,920	\$94	\$173	83%
Std. HDD	4,096	\$201	\$369	84%
Std. HDD Average				82%
High Perf. SSD	512	\$67	\$60	11%
High Perf. SSD	1,920	\$236	\$225	5%
High Perf. SSD	4,096	\$494	\$480	3%
High Perf. SSD Average				6%

The database average was sufficient for SSD disk drives (within about 5%), but the database average for HDD was severely overestimating.

c) Network Rates

The table below is a comparison of the commercial vendor database average rates to the Agency 1 rates for data transfer and operations requests. The transfer prices and “Write” operational requests are within an acceptable range, but the read requests have an unacceptable amount of error.

Table 23. Agency 1 Network - Model Results Comparison

Action	Agency 1 Average Monthly Rate	Commercial Vendor Avg Monthly rate	Absolute % Error
Transfer Pricing			
Transfer for 1,420 GB	\$141.55	\$138.89	2%
Operational Requests ⁺			
WRITE	\$0.030/GB	\$0.0368 /10K Requests	23%
READ	\$0.030/GB	\$0.0034 /10K Requests	89%

**The agency rate was given as a fixed of \$0.095/GB, so it was converted to TB and assumed to be the transfer rate for 0-1 TB*

**Operational Requests for the agency were given in terms of GB instead of request amounts*

iv. Agency 1 Conclusions

For the compute CERs, the average absolute average errors just over one standard deviation (on average 35% for our results versus 34% benchmark for one standard deviation) and the R^2 value >70% indicates that the CERs are borderline acceptable. As the number of CPUs and RAM gets smaller in a configuration, percent error generally gets larger. If Agency 1 had used the CERs developed in this paper to estimate the cost of their cloud requirement, the estimate may be under the actual amount needed. This may be because it is unknown how Agency 1 had calculated fixed rates for virtual machines based on ranges for CPUs and RAM memory.

The storage results are mixed. However, the overall vendor database average for HDD is 82% higher than the average for just AWS, indicating the vendor database for HDD contains rates that are much higher than the AWS rates. The AWS average values from the cloud vendor database for Object Storage varies from the Agency 1 model values by less than 6%. Due to this, the complete data set average for HDD disk drives may not be as valid as the average for specific vendors.

The results for data transfer are good. The rate for the agency appeared to be fixed, regardless of the number of terabytes transferred at \$97/TB/Month (\$0.095/GB). This value is very close to the vendor average value of \$96 for 0-1 TB and \$93.70 for 1- 10 TB, so the model calculation of \$138 for 1,420 GB was very close to the agency value of \$142. It is possible the fixed \$97/TB/Month rate for Agency 1 is an average for variable rates by number of terabytes transferred. But we did not have enough information from the agency model to make that assumption.

The basis for the operation request rates is incompatible and cannot be compared. We did not have sufficient information from the agency model to estimate the number of requests per GB used to establish their rate per GB. Nor did we know from the commercial vendor database how many GB were in 10,000 requests, although we did assume 1 GB was equal to 10,000 requests and did not come up with acceptable results for "Read" type operations, although the "Write" type was within an acceptable range. Therefore, it is not possible to reconcile the commercial database average rate approach with the fixed rates for the Agency 1 models.

3) Case 3: Agency 2

i. Agency Data

a) Compute Environment

The baseline environment at Agency 2 was planned to be replaced by AWS instances is shown below. Note how each replacement AWS instance is not an exact fit for the existing Agency 2 environment. The agency configurations include CPU + MEM + Block Storage components associated with their virtual machines, plus additional CPUs, Memory and Object Storage to support computing requirements. It was unclear how many virtual machines there were. The

replacement AWS instance configurations only include CPU + MEM + Block Storage components along with the amount of virtual machines for this specific configuration.

Table 24. Agency 2 Compute Configurations

Configuration	CPU's	Memory (GB RAM)	Block Storage (GB)	Total VMs	Added CPU's	Added GB RAM	Added Object Storage (GB)
Agency Config 1 (Large)	4	16	30		0	0	6,600
AWS Replacement: M1.xlarge	4	15	1,680	4			
Agency Config 2 (X-Large)	4	32	30		0	7.2	2,460
AWS Replacement: M2.2xlarge	4	34	850	3			
Agency Config 3 (X-Large)	4	32	30		80	728	33,000
AWS Replacement: M2.4xlarge	8	68	1,680	20			
Agency Config 4 (X-Large)	4	32	30		36	0	10,000
AWS Replacement: M3.2xlarge	8	30	30	9			

The prices for the AWS instance rates are shown below. These are 1 and 3 year reserved rates, based on 720 hours per month, for a Linux operating system. The 3 year discount was larger than the 1 year reserved rates, which was not logical. Therefore, the 3 year rates were disregarded, and only 1 year discount rates were compared.

Table 25. Agency 2 Compute Configuration Prices

Configuration	Number of VMs	System Cost – 1 Year Discount	System Cost – 3 Year Discount
AWS Replacement: M1.xlarge	4	\$6,772	\$25,576
AWS Replacement: M2.2xlarge	3	\$10,869	\$21,822
AWS Replacement: M2.4xlarge	20	\$146,400	\$295,420
AWS Replacement: M3.2xlarge	9	\$53,613	\$112,158

b) Storage Rates

The added object storage shown above was replaced with an AWS S3 storage instance, including 5 TB of disk capacity. Standard storage rates per Month for 5 TB were as follows: \$98.28/TB (\$0.11/ GB) for the first TB and \$112/TB (\$0.096/ GB) for the next 4 TB. The agency summed those prices for a monthly cost and then calculated a yearly cost based on that amount.

Table 26. Agency 2 Storage Pricing

Average per TB	For 1st TB	For Next 4 TB	Per Month	Per Year
\$ 98.28	\$ 112.00	\$ 392.00	\$ 504.00	\$ 6,048.00

c) Transfer Rates

WRITE and READ requests were included with the S3 instance, at an assumed level of 25 Million requests per month.

Table 27. Agency 2 Operational Request Pricing

Data Requests	Per 1,000 Requests	Price per Month
WRITE	\$ 0.0050	\$125
READ	\$ 0.0040	\$100

ii. Research Model Calculations

a) Compute CER

The model used Equation 5 in [Table 7](#) to estimate the 1 Year Reserved Hourly Rate, then multiplied the rate times 720 Hours per Month to get the equivalent Annual Rate per VM. The Annual Rate was then multiplied times the number of VMs to get the equivalent Annual Rate for the desired configuration. Two different compute configurations were modeled:

1. The number of CPUs, VMs, Memory, and Block Storage, plus the additional RAM and Object Storage in the current agency's configuration. This approach answers the question: what would be the estimated cost to replace the agency's current configuration as-is at the CER rates? This creates the desired comparison between the CER estimated cost to replace the current vendor configuration. Essentially, if the CER estimated cost for the complete current configuration is not significantly different from the estimated costs for the AWS replacement instances at their rates, then we would conclude the agency would have funded the same amount using the CER as was estimated for the AWS specific replacement instances. However, this configuration also has "Additional vCPUs" and "Additional vRAM" which could be added to the model. For example, for Agency Config 4 X-Large, add 36 additional vCPUs to the initial 4 to use 40

vCPUs in the CER. Overall, it was unclear how these additional vCPUs and RAM fit into these instances, so this approach was discarded.

Table 28. Agency 2 Compute – Configuration 1 Model Results

Current Agency Configurations	CER Calculated Rate per VM (1 Year Discount)	Number of VMs	CER Calculated Cost for the Agency Configuration
M1.Xlarge	\$ 1,153	4	\$ 4,611
M2.2xlarge	\$ 2,027	3	\$ 8,109
M2.4xlarge	\$ 30,592	20	\$ 142,709
M3.2xlarge	\$ 3,383	9	\$ 64,219

- The number of CPUs, VMs, Memory, and Block Storage to be provided by the AWS instance replacing the agency's current configuration ("AWS Replacement" rows in [Table 24](#)) were used in [Table 7](#) Eq 4. This approach answers the following question: what would be the estimated cost to replace the current configuration with the specific AWS configurations at the estimated CER rates? This creates a comparison between the commercial vendor Database Average Rates and the AWS specified rate researched by the agency for just the AWS configuration.

Table 29. Agency 2 Compute – Configuration 2 Model Results

AWS Specific Instance	CER Calculated Rate per VM (1 Year Discount)	Total VMs	CER Calculated Cost for the AWS Configuration
M1.Xlarge	\$ 1,527	4	\$ 6,109
M2.2xlarge	\$ 2,440	3	\$ 7,319
M2.4xlarge	\$ 4,817	20	\$ 96,333
M3.2xlarge	\$ 2,077	9	\$ 18,691

This comparison would be more logical if the AWS configurations were estimated with the CER using just the AWS instance specific rates from the database, and not the database average rates. The comparison would determine whether the AWS rates in the commercial vendor database were equal to the Agency 2 model AWS rates. However, this was not an objective of the research.

b) Storage Rates

First, we applied the commercial database average rates for Object Storage to the Agency 2 replacement configuration of 5 TB for Object Storage ([Table 9](#)). This essentially compares the cost of an AWS solution (S3 storage instance) to the database average for all commercial vendor storage solutions.

Table 30. Agency 2 Storage – Model Results

Comparison	Average per TB	For 1st TB	For Next 4 TB
Commercial Vendor DB	\$ 22.36	\$ 22.36	\$ 89.43

c) Network Rates

The only analysis possible was to compare the commercial vendor DB averages for Data Transfer and Operational Requests to the AWS specific rates (Table 9). The Agency model did not include a current rate for a current network configuration or bandwidth.

Table 31. Agency 2 Network – Model Results

Data Requests	Rate per 1,000 Requests	
	DB Avg	DB per Month
WRITE	\$ 0.0037	\$ 92
READ	\$ 0.0003	\$ 9

iii. Results Comparison

a) Compute CER

This is the cost for configuration 1, the AWS compared to the CER estimate for the current agency configuration with a 1 Year Discount:

Table 32. Agency 2 Compute - Model Results Comparison

AWS Specific Instance	CER Calculated Rate per VM (1 Year Discount)	Times Total VMs	CER Calculated Cost for the AWS Configuration	AWS Total Cost with 1 Year Discount	Delta (AWS - CER) for 1 Year Discount	% Error (AWS - CER)
M1.Xlarge	\$ 1,527	4	\$ 6,109	\$ 6,772	\$ 663	10%
M2.2xlarge	\$ 2,440	3	\$ 7,319	\$ 10,869	\$ 3,550	33%
M2.4xlarge	\$ 4,817	20	\$ 96,333	\$ 146,400	\$ 50,067	34%
M3.2xlarge	\$ 2,077	9	\$ 18,691	\$ 53,613	\$ 34,922	65%
					Absolute % Error	35%

The CER estimated cost for Compute with a 1 Year Discount is within 35% of the AWS costs for just the specific AWS compute instances (CPU + MEM + Block Storage).

b) Storage Rates

Cost for the AWS configuration using the commercial database average compared to the AWS rate:

Table 33. Agency 2 Object Storage - Model Results Comparison

Comparison	Average \$/TB	For 1st TB	For Next 4 TB	Per Month	Per Year	Delta (Agency – DB Mean)	% Error
Commercial Vendor DB	\$ 22.36	\$ 22.36	\$89.43	\$ 111.79	\$ 1,341.44	\$ 4,706.56	78%
Agency	\$ 98.28	\$ 112.00	\$ 392.00	\$ 504.00	\$6,048.00		

The rate used in the Agency 2 model is significantly higher than the commercial vendor database average rate by 78%.

c) Network Rates

The table below presents a comparison of the costs for Data Transfer and Operational requests. The agency values are generally much higher than the commercial database averages for all request actions.

Table 34. Agency 2 Requests - Model Results Comparison

Operational Requests	Rate per 1,000 Requests				
	Agency	Agency per Month	DB Avg	DB Avg per Month	Absolute % Error
WRITE	\$ 0.0050	\$ 125	\$0.0037	\$ 92	26%
READ	\$ 0.0040	\$ 100	\$0.0003	\$ 9	92%

iv. Agency 2 Conclusions

On average, the compute results were 35% lower than the Agency model, a borderline acceptable solution. The CER would have led to an organization underestimating their compute costs for budgeting purposes. An R^2 value was not calculated for only 4 data points.

As stated before, the agency monthly prices for object storage were nearly 80% higher than the database averages. In [Table 9](#), the AWS specific averages in the database are not very far from the database average used in the model (less than a 5% difference), so it is possible the agency model AWS storage is very different than the standard storage rates used in the model.

The prices for the data transfer and operational requests are difficult to compare. Assuming the level of 25 million requests applies to the current configuration and the AWS S3 solution, the only difference is in the rates used. The rates reported for the AWS S3 solution are significantly higher than the rates based on the commercial database averages for 6 vendors. The “Write”

requests are within an acceptable error, but the agency “Read” requests are very far off from the database averages. This cannot be reconciled. It is important to note that even the average Amazon “Read” requests across all regions in [Table 9](#) is \$0.0041 per 10,000 (or \$0.00041 per 1,000 requests). This is off by a magnitude of 10 of the agency model.

Overall, the compute rates for this agency are marginally acceptable. The storage and networking rates do not match this agency model.

4) Case 4: Agency 3 (Model 4)

i. Agency Data

a) Compute Environment

Agency 3 developed a complex workbook to evaluate the total cost of transitioning over 80 IT systems to a commercial cloud environment. This analysis examines only the estimated IaaS fees, based on detailed descriptions of the compute and storage requirements for each system in the cloud (VMs, CPUs, MEM, Block Storage).

The table below provides an extract of the data for 6 random systems in the agency model. The estimate included a transition scenario covering four fiscal years per system. No information was provided on the scenario per year (the percent of each system transferring each year), so this analysis assumes the annual average rate of transfer is the mean of the annual rates for each system.

Table 35. Agency 3 Compute - Data

System	CPUs	Memory (GB RAM)	Block Storage (GB)	VMs	Year 1	Year 2	Year 3	Year 4	Total
System 37	1	1	65	1	\$ 834	\$ 624	\$ 484	\$ 417	\$ 2,359
System 50	2	8	95	1	\$ 1,170	\$ 625	\$ 373	\$ 576	\$ 2,743
System 36	4	16	65	1	\$ 2,462	\$ 1,448	\$ 957	\$ 724	\$ 5,591
System 54	4	16	100	1	\$ 2,542	\$ 1,498	\$ 991	\$ 753	\$ 5,784
System 55	4	16	100	1	\$ 2,542	\$ 1,498	\$ 991	\$ 753	\$ 5,784
System 56	4	16	100	1	\$ 2,542	\$ 1,498	\$ 991	\$ 753	\$ 5,784

The final set of programs evaluated included only programs with a complete specification for the desired cloud configuration. Additionally, any programs that had specifications outside of the dataset range ([Table 4](#)) were also excluded. This reduced the dataset to 72 datapoints.

b) Storage and Network Rates

The agency model did not include any additional storage, nor did they include network costs in their models.

ii. Research Model Calculations

This agency model had different configurations of compute instances for specific vendors to one another spread over four years. Rates for four different vendors were included in the agency model, then compared to determine the most affordable cost. This analysis used the final agency rate as the comparison value.

The base rate CER for CPU + MEM + Block Storage was used to estimate the costs for each program transitioning to the cloud ([Table 7](#), Eq 4). This on-demand hourly rate per VM was then multiplied times by 720 hours per month and the number of VMs to arrive at the total annual cost for the new system. This is the annual cost for the complete system and was not spread at different amounts over four years like it was in the agency model. Thus, the CER annual rate times four years was compared to the agency total rate to evaluate the quality and fit of the model. Of course, this approach assumes that for each system that the spread of cost (and VM usage hours) over the four years was equal, but this of course is not the case.

iii. Results Comparisons

a) Compute Environment:

The table below shows the comparison of the CER total cost compared to the agency total cost for four years.

Table 36. Agency 3 – Compute Results Comparison

Goodness of Fit		Difference: Actual Cost - Estimated Cost	
Absolute % Error	30%	29 of 72 programs differ by less than + or - 15%	
R ²	0.91	48 of 72 programs differ by less than + or - 25%	
		42 Programs over estimated; 30 Underestimated	

In general, the CER tends to overestimate, but 48 of the 72 programs evaluated are within 25% (over or under) of the four-year agency estimate for the stated configuration.

Note, though, that the agency used IBM rates for their final results. In the commercial vendor database for all compute instance rates, the IBM rates in general presented the lowest average compared to all of the other rates in the vendor database. Therefore, the CER is probably skewed away from the IBM rates towards the average. As mentioned above, there was no indication of how many exact hours of virtual machine usage there was per year, so this analysis assumed of an equal spread of hours every year.

b) Agency 3 Conclusions

The R² measure of 0.91 is an acceptable fit, along with the average percent error of 30%. Looking at the overall range in the comparison, there is about a 65% chance that using the CERs

would have been within 25% of the agency estimate. If the agency used the CERs to estimate a budget close to the database mean for all vendors, they would have estimated more money than required to fund what is on average the lowest priced vendor in the database.

D. Overall Validation Conclusions

Our models/CERs were not perfect fits to the agency models because they were formed from a database of 6 different vendors across all regions of the US, while agency models were prices from one vendor and presumably from one region. The goal was to demonstrate that using the mostly commercial “vendor agnostic” approach presented in this paper would yield results just as valid as the agency model. This was done by comparing the results of each with the average absolute percent error. For situations with more than five virtual machine datapoints, Pearson’s R, and R² were also used in the analysis. Overall, the models had mixed results:

Table 37. Absolute Average Errors for the Compute CER in Validation and Comparison

Case	Case 1: Testing against random data in the dataset						Cases 2, 3, and 4: Comparison of Agency Models		
Abs Avg	Eq 1	Eq2	Eq3	Eq 4	Eq 5	Eq 6	A1 (Eq 1)	A2 (Eq 5)	A3 (Eq 4)
% Error	37%	31%	27%	20%	23%	22%	35%	35%	30%
								Average	29%

The average absolute errors for compute prices for each agency model and the random sampling comparison (Case 1) were, a majority of times, below one standard deviation ($\pm 34\%$). In the handful of times this wasn’t true, they were no more than 37% which is very close. This is a more acceptable error for monthly or yearly prices (all agency models), but not as acceptable for hourly price comparisons (Case 1).

The error associated with compute pricing in the validation above could come from many sources. It could be demonstrating that cloud pricing varies among different providers. It is also unknown what fiscal year these agency models are from, but the data collected for this paper was from the summer and Fall of 2019. In addition, all agency models were also difficult to match to the ones developed in this paper. Agency 1 had a range of specifications attributed to a single price. The second agency had one type of configuration version that was impossible to reconcile with our data because of missing information. For the last agency, it was assumed that there was an even hourly usage for the virtual machines across four years for each system.

Storage and network related models were not tested thoroughly among the three agencies. Transfer pricing and SSD disk pricing in the vendor agnostic model were very close to the agency model for the first agency. For that same agency, HDD prices were far off from the database average, but were shown to be closer to the AWS only prices. For two agencies,

“Write” request operations were within acceptable range of error (23% and 26%) but “Read” request operations were about 90% off.

Errors for these types of models can also come from many sources. The dataset has smaller amounts of data which may have had variation. Furthermore, the price per amount of disk space could vary widely for a single region, while the price per GB for IBM was fixed. In addition, there was no opportunity to test the file storage or snapshot models because they were not included in any agency models.

5. Challenges and Next Steps

A. Challenges

1) Data collection

Throughout the research, one of the largest challenges was collecting data by hand instead of in an automated fashion. Data collection took a lot more time than expected which limited the time taken to explore building pricing models to represent other cloud services (such as database storage and management).

Some virtual machines can have additional disks attached to them. Due to a lack of time, disk properties were not included in VM data for the following vendors: Oracle, Alibaba, IBM and Google. In the cases where disk properties were collected, the disk storage size had to be matched to the virtual machine in the dataset. Overall, the ETL process used was ad hoc and also across multiple programs (e.g. RapidMiner and Excel). In hindsight, the programs being used were not necessarily the best for ETL, but they were ones that the authors had experience in using.

2) Normalization

Each provider sets up their pricing structures differently, so the data had to be amalgamated into one format to best represent the information from all six providers. Nuances between pricing styles and definitions may have been lost; due to the interest of time, not all provider documentation could be read thoroughly.

When collecting Virtual Machine data, both Compute and Kubernetes data types were collected. After collecting both, it was realized that the pricing for both is identical. However, it was determined that the identical data didn’t affect the CER dramatically enough to reconsider redoing validation. In future versions of these analysis the datapoints will be removed.

3) Discarded research

Earlier in the research process, it was thought that if pricing data across different locations followed a pattern (or one could be approximated) then a model could be built based off of data from one location. Then, “location factor multipliers” could be created to solve for prices from different locations. This could make future updates to the models and CERs easier. However, when this was tested with at least one vendor there seemed to be no consistent pattern to pricing taken across different locations.

B. Next steps

There are many possibilities for future research. One task that is critical is to mitigate error by adding uncertainty to the model. For the CERs, if uncertainty was selected at the median confidence level, the value would mitigate the probability of underestimating as compared to specific vendor models.

There were other research topics left out due to a lack of time. There was international cloud services pricing data that was never analyzed. In addition, more investigation needs to be done on file storage networking costs and pricing structures. There are more cloud services and cloud related costs left out of this study. This includes the costs of modifying and migrating data, along with costs of refactoring software for re-hosting.

Automating more data collection and research would be more beneficial for future iterations of the project. A more robust and organized ETL process could be used to transform the data and prepare it for analysis more effectively. A language like R may streamline the analysis process [25].

6. Conclusion

This paper presents CERs and models to predict pricing for cloud compute and storage pricing. These models were created from over 28,000 open source datapoints. After the CERs and models were created they were compared to a random subset of the original dataset and also to three models from federal agencies. If there was no statistically significant difference, then it could be claimed that the CERs/models could be used in the place of the agency calculators.

The results show that the CERs for compute pricing are borderline acceptable to their agency counterparts and the random sampling of data. These tests had absolute average errors of 20 – 37%, with an average of about 30%. These CER estimates are correlated well with their agency model counterparts. The results for the storage models were mixed: for example, the average price per GB for an additional SSD disk performed well, while the HDD counterpart performed poorly.

More testing needs to be done on these models, and an uncertainty analysis should be carried out. In the meantime, users are able to take this open source data to focus on the datapoints that are most relevant to their own studies. However, in less than a year the data in this model will need to be updated due to the dynamic nature of cloud pricing. Finally, more research should be carried out on additional cloud costs such as data migration.

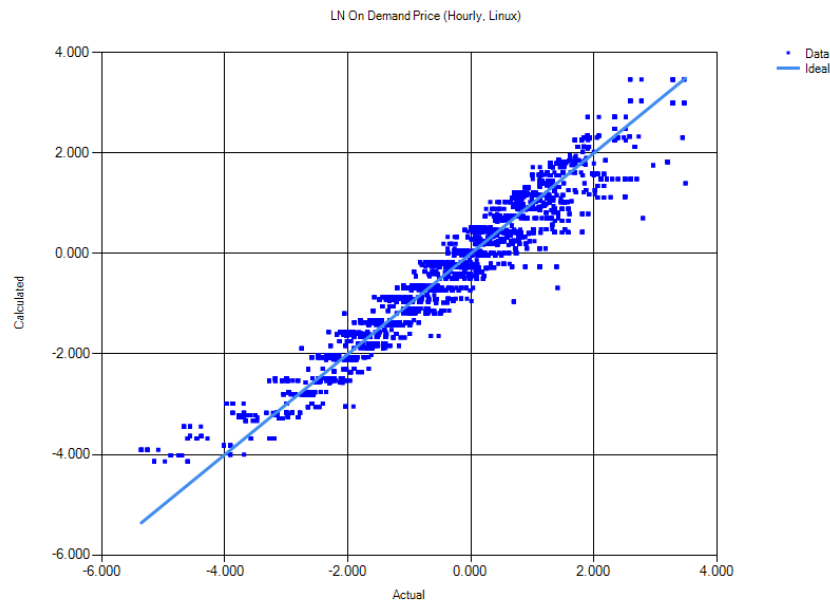
7. Bibliography

- [1] R. Mabe and D. Harper, "Using Predictive Analytics and Open Source Data to Estimate IT and Cloud Related Costs for Government IT Systems," May 2019. [Online]. Available: <http://www.iceaaonline.com/ready/wp-content/uploads/2019/06/CC03-Predictive-Analytic-Estimates-of-Cloud-Costs-Mabe.pdf>. [Accessed February 2020].
- [2] Amazon, "AWS Home Page," [Online]. Available: https://aws.amazon.com/?nc2=h_lg. [Accessed February 2020].
- [3] Microsoft, "Microsoft Azure Homepage," [Online]. Available: <https://azure.microsoft.com/en-us/>. [Accessed February 2020].
- [4] Alibaba, "Alibaba Cloud," [Online]. Available: <https://us.alibabacloud.com/>. [Accessed February 2020].
- [5] Google, "Google Cloud Homepage," [Online]. Available: <https://cloud.google.com/>. [Accessed February 2020].
- [6] IBM, "IBM Cloud," [Online]. Available: <https://www.ibm.com/cloud>. [Accessed February 2020].
- [7] Oracle, "Cloud Infrastructure," [Online]. Available: <https://www.oracle.com/cloud/>. [Accessed February 2020].
- [8] Office of the Federal Chief Information Officer, "M-16-19 - Data Center Optimization Initiative," 1 August 2016. [Online]. Available: <https://policy.cio.gov/dcoi/>. [Accessed February 2020].
- [9] S. Kent, "MEMORANDUM FOR CHIEF INFORMATION OFFICERS OF EXECUTIVE," 25 June 2019. [Online]. Available: <https://www.whitehouse.gov/wp-content/uploads/2019/06/M-19-19-Data-Centers.pdf>. [Accessed February 2020].
- [10] Wikipedia, "Main Page," [Online]. Available: https://en.wikipedia.org/wiki/Main_Page. [Accessed February 2020].
- [11] G. Portella, G. Rodrigues, E. Nakano and A. C. Melo, "Statistical Analysis of Amazon EC2 Cloud Pricing Models," *Concurrency and Computation Practice and Experience*, pp. 1-16, 2018.
- [12] J. Smirnoff and H. Souiri, "Forecasting Future Amazon Web Services Pricing," in *ITCAST Symposium 2019 Program*, Crystal City, 2019.
- [13] M. Zhang, R. Ranjan, S. Nepal, M. Menzel and A. Haller, "A Declarative Recommender System for Cloud Infrastructure Services Selection," October 2012. [Online]. Available: https://www.researchgate.net/publication/232062949_A_Declarative_Recommender_System_for_Cloud_Infrastructure_ServicesSelection. [Accessed Feb 2020].

- [14] R. Stephens, "IaaS Pricing Patterns and Trends 2018," RedMonk, 13 July 2018. [Online]. Available: <https://redmonk.com/rstephens/2018/07/13/iaas-pricing-patterns-and-trends-2018/>. [Accessed February 2020].
- [15] R. Stephens, "IaaS Pricing Patterns and Trends 2019," RedMonk, 1 August 2019. [Online]. Available: <https://redmonk.com/rstephens/2019/08/01/iaas-pricing-patterns-and-trends-2019/>. [Accessed February 2020].
- [16] E. Lu, "Cloud Costs Aren't Dropping Dramatically," Kapwing, 2 January 2018. [Online]. Available: <https://www.kapwing.com/blog/cloud-costs-arent-actually-dropping-dramatically/>. [Accessed February 2020].
- [17] Banzai Cloud, "Cloud Instance Type Details," [Online]. Available: <https://banzaicloud.com/cloudinfo/>. [Accessed February 2020].
- [18] G. Heaton, "EC2Instances.info: Easy Amazon EC2 Instance Comparison," [Online]. Available: https://www.ec2instances.info/?cost_duration=monthly. [Accessed February 2020].
- [19] "cURL:// Command Line Tool and Library," [Online]. Available: <https://curl.haxx.se/>. [Accessed February 2020].
- [20] RapidMiner, Inc, [Online]. Available: <https://rapidminer.com/>.
- [21] Octopus Data Inc., [Online]. Available: <https://www.octoparse.com/>. [Accessed February 2020].
- [22] C. Cuiule and G. Noll, "Big Data" Analytics in Operations Research," May 2019. [Online]. Available: <http://www.iceaaonline.com/ready/wp-content/uploads/2019/06/DM01-Paper-Big-Data-Analytics-in-Operations-Research-Cuiule.pdf>. [Accessed February 2020].
- [23] C. Harvey, "Cloud Storage Pricing: Top Vendors Price Comparison," Enterprise Storage, 26 April 2018. [Online]. Available: <https://www.enterprisestorageforum.com/cloud-storage/cloud-storage-pricing.html>. [Accessed February 2020].
- [24] A. K, "Requests Prices in Azure, Google and AWS Compared," 28 September 2016. [Online]. Available: <https://www.msp360.com/resources/blog/requests-and-data-transfer-prices-in-azure-google-and-aws-compared/>. [Accessed February 2020].
- [25] "The R Project for Statistical Computing," The R Foundation, [Online]. Available: <https://www.r-project.org/>. [Accessed February 2020].
- [26] S. Donovan, "Memorandum For Heads of Executive Departments and Agencies," 10 June 2015. [Online]. Available: <https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2015/m-15-14.pdf>. [Accessed February 2020].

8. Attachment A: Detailed Statistical Results for Each CER

Equation 1: $\text{Rate} = \exp(0.662 * \text{LN}[\text{RAM}] + 0.337 * \text{LN}[\text{vCPUs}] - 3.680)$



Statistics

>>

Regression Table

Statistics	Value
Multiple R	0.967
R Square	0.936
Adjusted R Square	0.936
Standard Error	0.392
Observations	6391.000

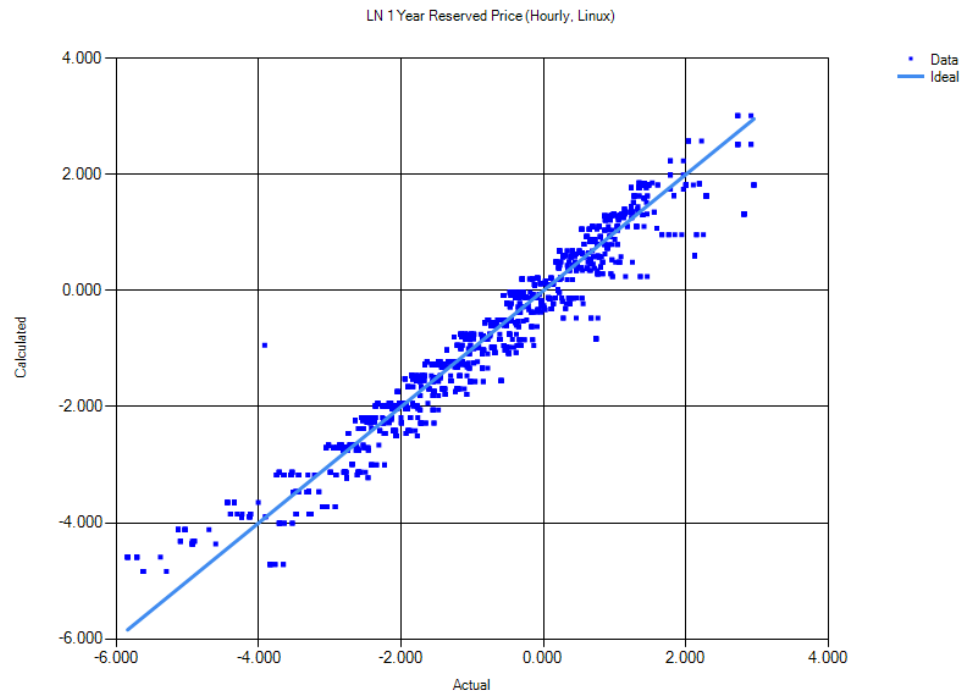
Anova Table

Source	Sum of Squares	Degrees of Freedom	Mean Squares	P-Value	F-Statistic
Regression	14236.461	2	7118.231	0.000	46441.322
Error	979.112	6388	0.153		
Total	15215.573	6390	2.381		

Coefficients Table

Name	Coefficient Value	Std. Error	t Test	P-Value	Upper Confident...	Lower Confident...
LN CPUs	0.337	0.008	43.610	0.000	0.353	0.322
LN Memory (RA...	0.662	0.006	116.218	0.000	0.673	0.651
Intercept	-3.680	0.012	-308.062	0.000	-3.656	-3.703

Equation 2: $1 \text{ Yr Disc. Rate} = \exp(0.679 * \text{LN}[\text{RAM}] + 0.356 * \text{LN}[\text{vCPUs}] - 4.367)$



Statistics

Regression Table

Statistics	Value
Multiple R	0.971
R Square	0.943
Adjusted R Square	0.943
Standard Error	0.396
Observations	2215.000

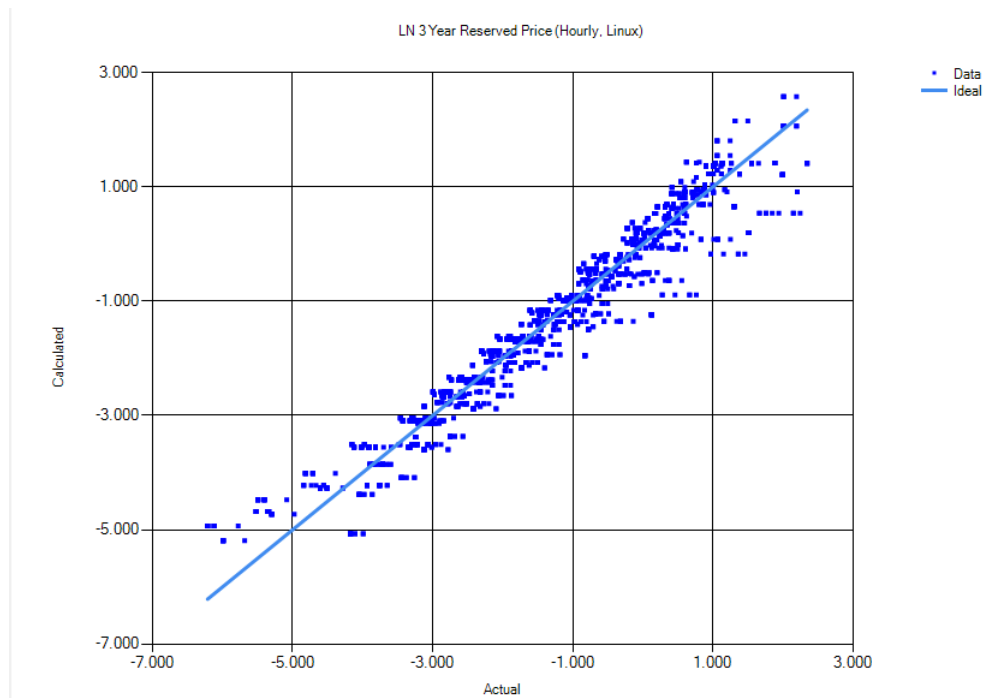
Anova Table

Source	Sum of Squares	Degrees of Freedom	Mean Squares	P-Value	F-Statistic
Regression	5766.477	2	2883.238	0.000	18420.216
Error	346.235	2212	0.157		
Total	6112.712	2214	2.761		

Coefficients Table

Name	Coefficient Value	Std. Error	t Test	P-Value	Upper Confident...	Lower Confident...
LN CPUs	0.356	0.013	27.727	1.754E-145	0.381	0.331
LN Memory (RA...	0.679	0.010	69.999	0.000	0.698	0.660
Intercept	-4.367	0.021	-210.747	0.000	-4.326	-4.407

Equation 3: 3 Yr Disc. Rate = $\exp(0.663 * \text{LN}[\text{RAM}] + 0.370 * \text{LN}[\text{vCPUs}] - 4.731)$



Statistics

Regression Table

Statistics	Value
Multiple R	0.969
R Square	0.939
Adjusted R Square	0.939
Standard Error	0.409
Observations	2215.000

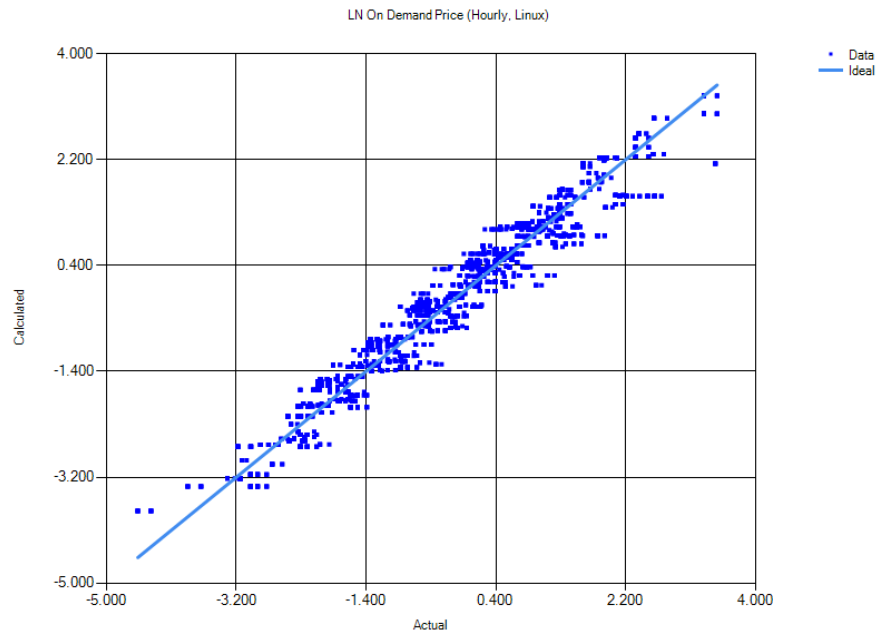
Anova Table

Source	Sum of Squares	Degrees of Freedom	Mean Squares	P-Value	F-Statistic
Regression	5687.908	2	2843.954	0.000	17026.242
Error	369.478	2212	0.167		
Total	6057.386	2214	2.736		

Coefficients Table

Name	Coefficient Value	Std. Error	t Test	P-Value	Upper Confident...	Lower Confident...
LN CPUs	0.370	0.013	27.935	2.416E-147	0.396	0.344
LN Memory (RA...	0.663	0.010	66.094	0.000	0.682	0.643
Intercept	-4.731	0.021	-221.016	0.000	-4.689	-4.773

Equation 4: $\text{Rate} = \exp(0.602 * \text{LN}[\text{RAM}] + 0.224 * \text{LN}[\text{vCPUs}] + 0.146 * \text{LN}[\text{Local Disk}] - 4.029)$



Statistics



Regression Table

Statistics	Value
Multiple R	0.974
R Square	0.949
Adjusted R Square	0.949
Standard Error	0.334
Observations	2884.000

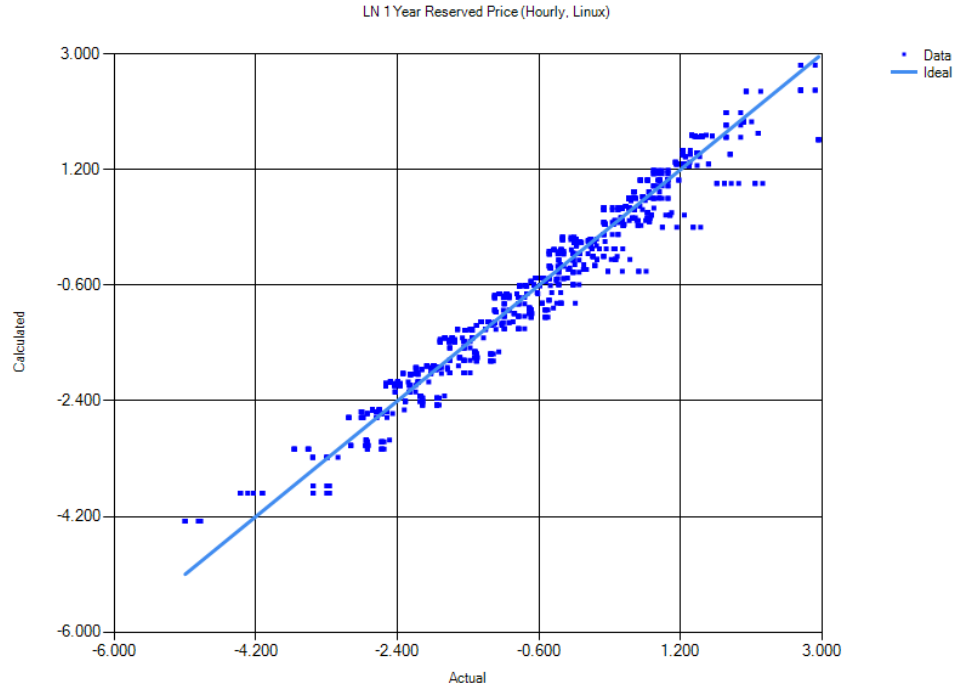
Anova Table

Source	Sum of Squares	Degrees of Freedom	Mean Squares	P-Value	F-Statistic
Regression	6032.496	3	2010.832	0.000	18035.161
Error	321.106	2880	0.111		
Total	6353.602	2883	2.204		

Coefficients Table

Name	Coefficient Value	Std. Error	t Test	P-Value	Upper Confident Limit	Lower Confident Limit
LN CPUs	0.224	0.010	21.988	3.472E-099	0.244	0.204
LN Local Disk (GB)	0.146	0.006	24.660	5.582E-122	0.158	0.135
LN Memory (RAM)...	0.602	0.009	70.683	0.000	0.618	0.585
Intercept	-4.029	0.021	-192.278	0.000	-3.988	-4.070

Equation 5: 1 Yr Disc. Rate = $\exp(0.630 * \text{LN}[\text{RAM}] + 0.278 * \text{LN}[\text{vCPUs}] + 0.080 * \text{LN}[\text{Local Disk}] - 4.433)$



Statistics

Regression Table

Statistics	Value
Multiple R	0.979
R Square	0.958
Adjusted R Square	0.958
Standard Error	0.304
Observations	1161.000

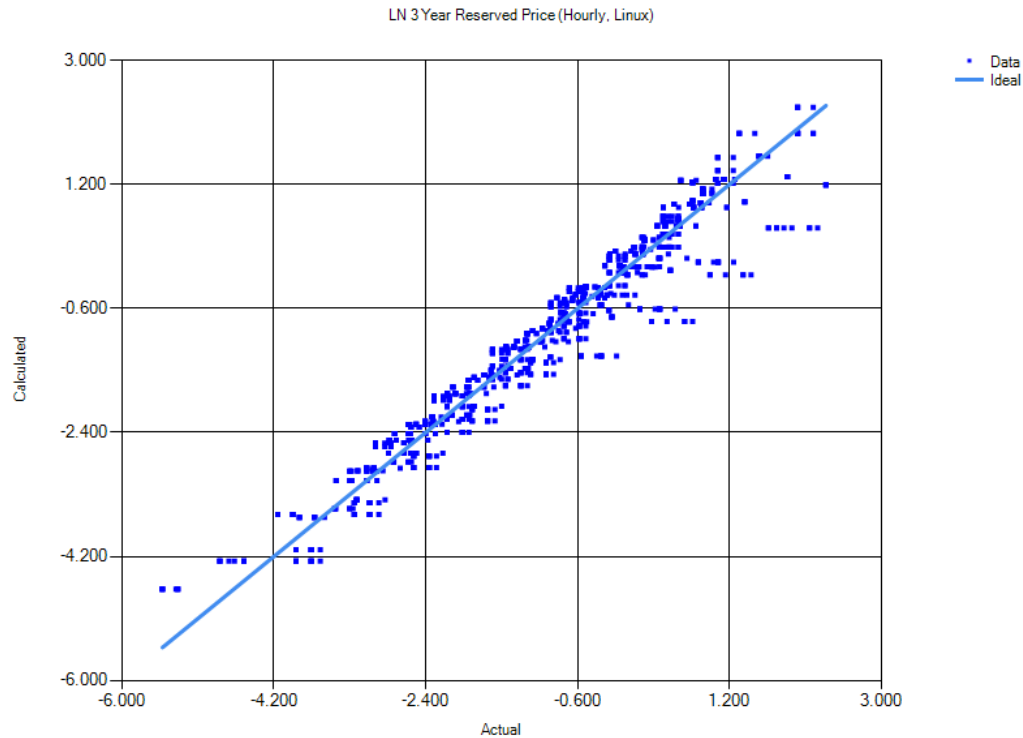
Anova Table

Source	Sum of Squares	Degrees of Freedom	Mean Squares	P-Value	F-Statistic
Regression	2453.402	3	817.801	0.000	8850.802
Error	106.905	1157	0.092		
Total	2560.307	1160	2.207		

Coefficients Table

Name	Coefficient Value	Std. Error	t Test	P-Value	Upper Confident Limit	Lower Confident Limit
LN CPUs	0.278	0.015	19.044	1.505E-070	0.307	0.249
LN Local Disk (GB)	0.080	0.009	9.181	1.912E-019	0.097	0.063
LN Memory (RAM...)	0.630	0.012	51.913	1.848E-304	0.653	0.606
Intercept	-4.433	0.033	-135.922	0.000	-4.369	-4.497

Equation 6: 3 Yr Disc. Rate = $\exp(0.588 * \text{LN}[\text{RAM}] + 0.270 * \text{LN}[\text{vCPUs}] + 0.120 * \text{LN}[\text{Local Disk}] - 4.891)$



Statistics

»

Regression Table

Statistics	Value
Multiple R	0.971
R Square	0.944
Adjusted R Square	0.944
Standard Error	0.353
Observations	1161.000

Anova Table

Source	Sum of Squares	Degrees of Freedom	Mean Squares	P-Value	F-Statistic
Regression	2417.265	3	805.755	0.000	6472.254
Error	144.039	1157	0.124		
Total	2561.304	1160	2.208		

Coefficients Table

Name	Coefficient Value	Std. Error	t Test	P-Value	Upper Confident Limit	Lower Confident Limit
LN CPUs	0.270	0.017	15.936	7.661E-052	0.303	0.237
LN Local Disk (GB)	0.120	0.010	11.803	1.940E-030	0.140	0.100
LN Memory (RAM) ...	0.588	0.014	41.803	1.715E-233	0.616	0.561
Intercept	-4.891	0.038	-129.192	0.000	-4.817	-4.965