WEEK 2 PAPER

Large-Scale Data Storage Systems – DATA-5400 | Spring 2020

Christina Morgenstern

Scenario 1: Start-up Company „CancerBytes“

My start-up company deals with patient data and tries to find patterns in genetic information to inform decisions on treatment options for cancer. We have just formed our company which comprises of three people and our shared office space is located in London, UK. Our infrastructure is solely based on our laptops, but the company would like to harness cloud services to be able to deal with all the data that has become unmanageable on our computers as well as to run our cancer prediction models at a reasonable speed.

Since our background is diverse with two Python developers and me as a biomedical data scientist, I would suggest choosing the Platform as a Service (PaaS) option with deployment in a hybrid cloud. The PaaS services gives us the freedom to write our own code and establish pipelines without being limited by preset apps. The hybrid cloud lets us both store sensitive patient information as well as use services of the public cloud when dealing with email accounts or distributing shareable data to the community.

Upon discussion of suitable PaaS service providers we decided on Heroku, a platform as a service based on a managed container system and with integrated data services for deploying and running modern apps (https://www.heroku.com). Benefits of Heroku include the Developer Experience which lets us focus on creating and delivering apps without paying attention to servers or infrastructure. Also, Heroku allows us to deploy code using our main programming language Python but it is flexible for many other languages too. Heroku´s data services allow apps to be customized with different add-ons that can provide further functionality to our apps such as data stores and monitoring. We also get good support from Heroku through application metrics, threshold alerting and other features that help with troubleshooting. Lastly, Heroku offers a secure platform for our patient data.

In order to select on a pricing plan, we visit the pricing calculator on Heroku (https://www.heroku.com/pricing#dynos). We decided on a Professional plan deploying Performance and Standard dynos in combination. Dynos are the heart of the Heroku platform and represent isolated, virtualized Linux containers that are designed to execute user-specific code. With one Performance L and two Standard 2X dynos the monthly costs would be 600$ (see Fig. 1 for calculation). We chose the highest performance dyno with 14GB RAM in combination with two standard dynos because the combination allows us to run our cancer prediction model with high performance.

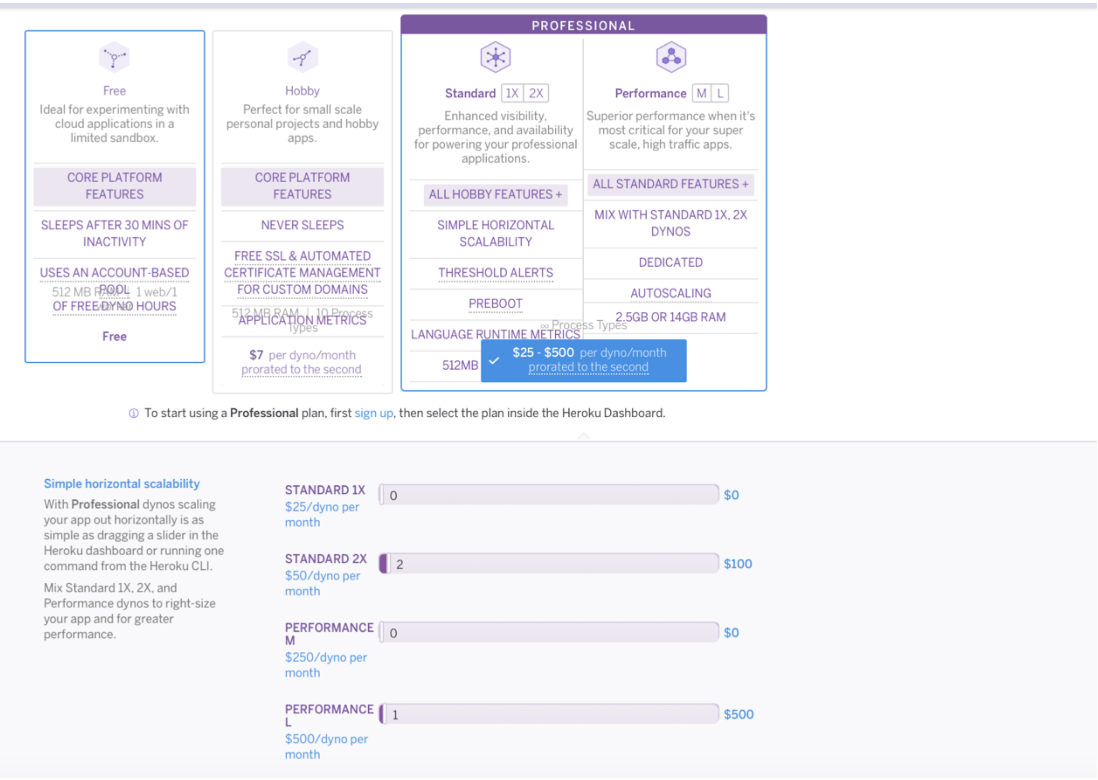
When adding storage of Standard 4 – 30 GB RAM, 750 GB storage 500 connections, the estimated monthly costs increase by 750$ to 1350$.

Heroku Redis further allows to scale for high speed RAM available for storing and caching data. We choose the Premium 0 – 50 MB RAM, 40 connections option.

Adding Apache Kafka Basicplan (Standard 0 – 4 GB capacity for 100$) will offer multi-tenant Kafka for development, testing and low volume production event streams.

In total the costs add up to 1465$ per month. Additionally, Heroku comes with free Standard Support which we might come back to if needed. We are happy with the chosen options and can afford Heroku at least for the coming months as funding has been secured.

Selecting the Dynos:



Free
Ideal for experimenting with cloud applications in a limited sandbox.

- CORE PLATFORM FEATURES
- SLEEPS AFTER 30 MINS OF INACTIVITY
- USES AN ACCOUNT-BASED POOL OF FREE DYNO HOURS
- 512 MB RAM | 1 web/1 worker

Free

Hobby
Perfect for small scale personal projects and hobby apps.

- CORE PLATFORM FEATURES
- NEVER SLEEPS
- FREE SSL & AUTOMATED CERTIFICATE MANAGEMENT FOR CUSTOM DOMAINS
- APPLICATION METRICS
- 512 MB RAM | 10 Process Types

$7 per dyno/month
prorated to the second

PROFESSIONAL

Standard 1X 2X
Enhanced visibility, performance, and availability for powering your professional applications.

- ALL HOBBY FEATURES +
- SIMPLE HORIZONTAL SCALABILITY
- THRESHOLD ALERTS
- PREBOOT
- LANGUAGE RUNTIME METRICS
- 512MB

Performance M L
Superior performance when it's most critical for your super scale, high traffic apps.

- ALL STANDARD FEATURES +
- MIX WITH STANDARD 1X, 2X DYNOS
- DEDICATED
- AUTOSCALING
- 2.5GB OR 14GB RAM
- ∞ Process Types

✓ $25 - $500 per dyno/month
prorated to the second

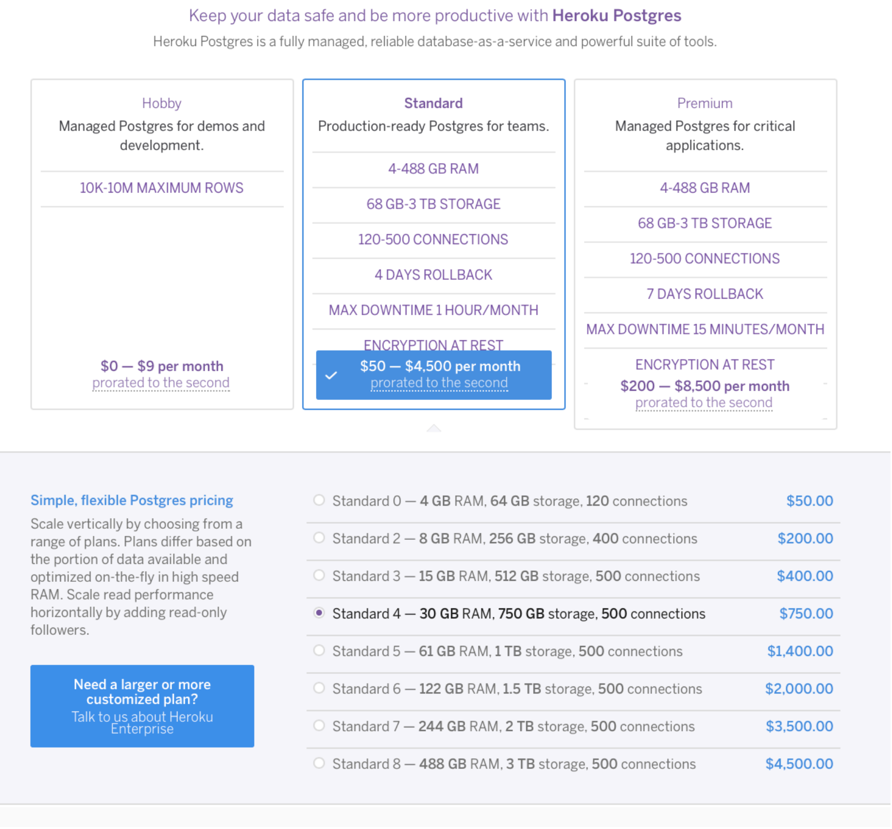
ⓘ To start using a **Professional** plan, first sign up, then select the plan inside the Heroku Dashboard.

Simple horizontal scalability

With **Professional** dynos scaling your app out horizontally is as simple as dragging a slider in the Heroku dashboard or running one command from the Heroku CLI.

Mix Standard 1X, 2X, and Performance dynos to right-size your app and for greater performance.

STANDARD 1X $25/dyno per month	0	$0
STANDARD 2X $50/dyno per month	2	$100
PERFORMANCE M $250/dyno per month	0	$0
PERFORMANCE L $500/dyno per month	1	$500



Keep your data safe and be more productive with **Heroku Postgres**

Heroku Postgres is a fully managed, reliable database-as-a-service and powerful suite of tools.

Hobby
Managed Postgres for demos and development.

- 10K-10M MAXIMUM ROWS

$0 — $9 per month
prorated to the second

Standard
Production-ready Postgres for teams.

- 4-488 GB RAM
- 68 GB-3 TB STORAGE
- 120-500 CONNECTIONS
- 4 DAYS ROLLBACK
- MAX DOWNTIME 1 HOUR/MONTH
- ENCRYPTION AT REST

✓ $50 — $4,500 per month
prorated to the second

Premium
Managed Postgres for critical applications.

- 4-488 GB RAM
- 68 GB-3 TB STORAGE
- 120-500 CONNECTIONS
- 7 DAYS ROLLBACK
- MAX DOWNTIME 15 MINUTES/MONTH
- ENCRYPTION AT REST

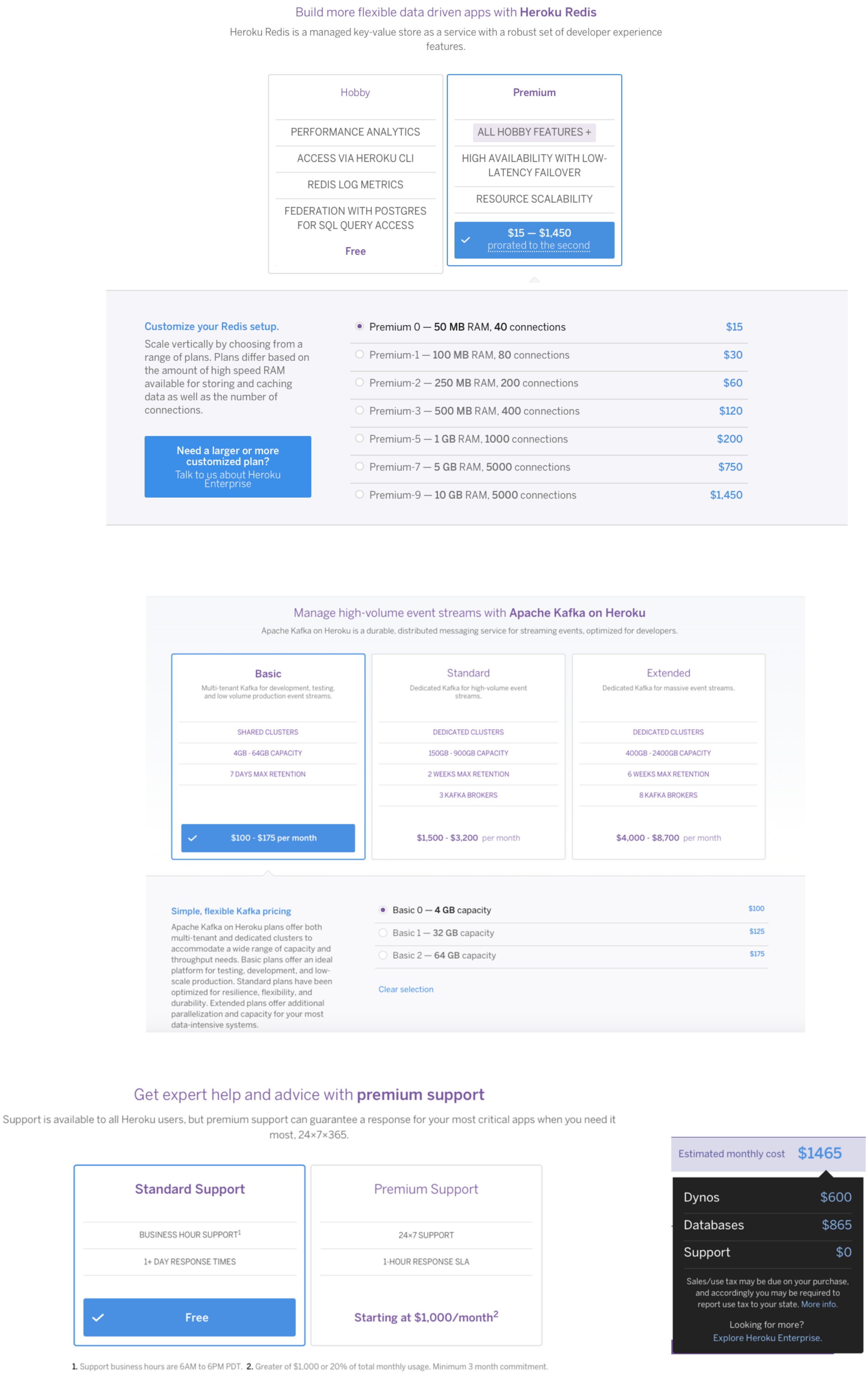
$200 — $8,500 per month
prorated to the second

Simple, flexible Postgres pricing

Scale vertically by choosing from a range of plans. Plans differ based on the portion of data available and optimized on-the-fly in high speed RAM. Scale read performance horizontally by adding read-only followers.

Need a larger or more customized plan?
Talk to us about Heroku Enterprise

○ Standard 0 — **4 GB** RAM, **64 GB** storage, **120** connections	$50.00
○ Standard 2 — **8 GB** RAM, **256 GB** storage, **400** connections	$200.00
○ Standard 3 — **15 GB** RAM, **512 GB** storage, **500** connections	$400.00
● Standard 4 — **30 GB** RAM, **750 GB** storage, **500** connections	$750.00
○ Standard 5 — **61 GB** RAM, **1 TB** storage, **500** connections	$1,400.00
○ Standard 6 — **122 GB** RAM, **1.5 TB** storage, **500** connections	$2,000.00
○ Standard 7 — **244 GB** RAM, **2 TB** storage, **500** connections	$3,500.00
○ Standard 8 — **488 GB** RAM, **3 TB** storage, **500** connections	$4,500.00



Build more flexible data driven apps with **Heroku Redis**

Heroku Redis is a managed key-value store as a service with a robust set of developer experience features.

Hobby	Premium
PERFORMANCE ANALYTICS	ALL HOBBY FEATURES +
ACCESS VIA HEROKU CLI	HIGH AVAILABILITY WITH LOW-LATENCY FAILOVER
REDIS LOG METRICS	RESOURCE SCALABILITY
FEDERATION WITH POSTGRES FOR SQL QUERY ACCESS	✓ $15 — $1,450 prorated to the second
Free	

Customize your Redis setup.

Scale vertically by choosing from a range of plans. Plans differ based on the amount of high speed RAM available for storing and caching data as well as the number of connections.

Need a larger or more customized plan?
Talk to us about Heroku Enterprise

● Premium 0 — **50 MB** RAM, **40** connections	$15
○ Premium-1 — **100 MB** RAM, **80** connections	$30
○ Premium-2 — **250 MB** RAM, **200** connections	$60
○ Premium-3 — **500 MB** RAM, **400** connections	$120
○ Premium-5 — **1 GB** RAM, **1000** connections	$200
○ Premium-7 — **5 GB** RAM, **5000** connections	$750
○ Premium-9 — **10 GB** RAM, **5000** connections	$1,450

Manage high-volume event streams with **Apache Kafka on Heroku**

Apache Kafka on Heroku is a durable, distributed messaging service for streaming events, optimized for developers.

Basic	Standard	Extended
Multi-tenant Kafka for development, testing, and low volume production event streams.	Dedicated Kafka for high-volume event streams.	Dedicated Kafka for massive event streams.
SHARED CLUSTERS	DEDICATED CLUSTERS	DEDICATED CLUSTERS
4GB - 64GB CAPACITY	150GB - 900GB CAPACITY	400GB - 2400GB CAPACITY
7 DAYS MAX RETENTION	2 WEEKS MAX RETENTION	6 WEEKS MAX RETENTION
	3 KAFKA BROKERS	8 KAFKA BROKERS
✓ $100 - $175 per month	$1,500 - $3,200 per month	$4,000 - $8,700 per month

Simple, flexible Kafka pricing

Apache Kafka on Heroku plans offer both multi-tenant and dedicated clusters to accommodate a wide range of capacity and throughput needs. Basic plans offer an ideal platform for testing, development, and low-scale production. Standard plans have been optimized for resilience, flexibility, and durability. Extended plans offer additional parallelization and capacity for your most data-intensive systems.

● Basic 0 — **4 GB** capacity	$100
○ Basic 1 — **32 GB** capacity	$125
○ Basic 2 — **64 GB** capacity	$175

Clear selection

Get expert help and advice with **premium support**

Support is available to all Heroku users, but premium support can guarantee a response for your most critical apps when you need it most, 24×7×365.

Standard Support	Premium Support
BUSINESS HOUR SUPPORT[1]	24×7 SUPPORT
1+ DAY RESPONSE TIMES	1-HOUR RESPONSE SLA
✓ Free	Starting at $1,000/month[2]

[1] Support business hours are 6AM to 6PM PDT. [2] Greater of $1,000 or 20% of total monthly usage. Minimum 3 month commitment.

Estimated monthly cost $1465

Dynos	$600
Databases	$865
Support	$0

Sales/use tax may be due on your purchase, and accordingly you may be required to report use tax to your state. More info.

Looking for more?
Explore Heroku Enterprise.

Figure 1. Screenshots demonstrating the selected purchasing options on Heroko.

Scenario 2: Large company

I am a Data Scientist within a large pharmaceutical company and data is at our heart in every area of the business. At our headquarter in Cambridge we have approx. 1000 employees and we have subsidiaries in more than 70 countries worldwide leading to a total of 30.000 employees. In the past the IT department in the company has operated using our huge data center. Planning, purchasing, installing and configuration of resources was the sole responsibility of the IT team. However, this solution is not practical anymore and according to our accountant team doesn´t pay off anymore. There are many under- and over utilized resources, temporarily restrictions in CPU and memory and the self-repeating claim-buy-install software life cycles. Thus, the management has decided to switch to a cloud computing solution to provide a more adequate, flexible and cost-effective solution to managing our data. Together with the IT department they decided to implement an Infrastructure as a Service (IaaS) based service to provide computing, networking and storage resources on demand which is provisioned and managed over the internet. Our IT team will still be responsible for configuring and managing each virtual resource but issues like servers, storage and networking will be serviced by the provider. The mode of deployment should be a hybrid cloud which can help integrate our on-premises data centers with cloud solutions.

Making use of an IaaS has the advantage that the business can be more agile, meaning that with the help of the IT team workloads can be better monitored, data accessed on demand and capacity increased according to needs. Security is also an issue. Since all sensitive data is stored and handled internally and not available to the general public, this option implies a higher level of security and privacy.

The company made the decision to use Amazon Web Services (AWS) as on-demand cloud computing platform. AWS were the pioneers in the field and it represents in essence a huge collection of IT infrastructures (servers, storage…) located in multiple data centers around the world offering more than 175 different features and services including compute, storage, databases, analytics, networking, mobile, developer tools, management tools, IoT, security and enterprise applications. AWS promises that these services help businesses to move faster with lower IT costs.

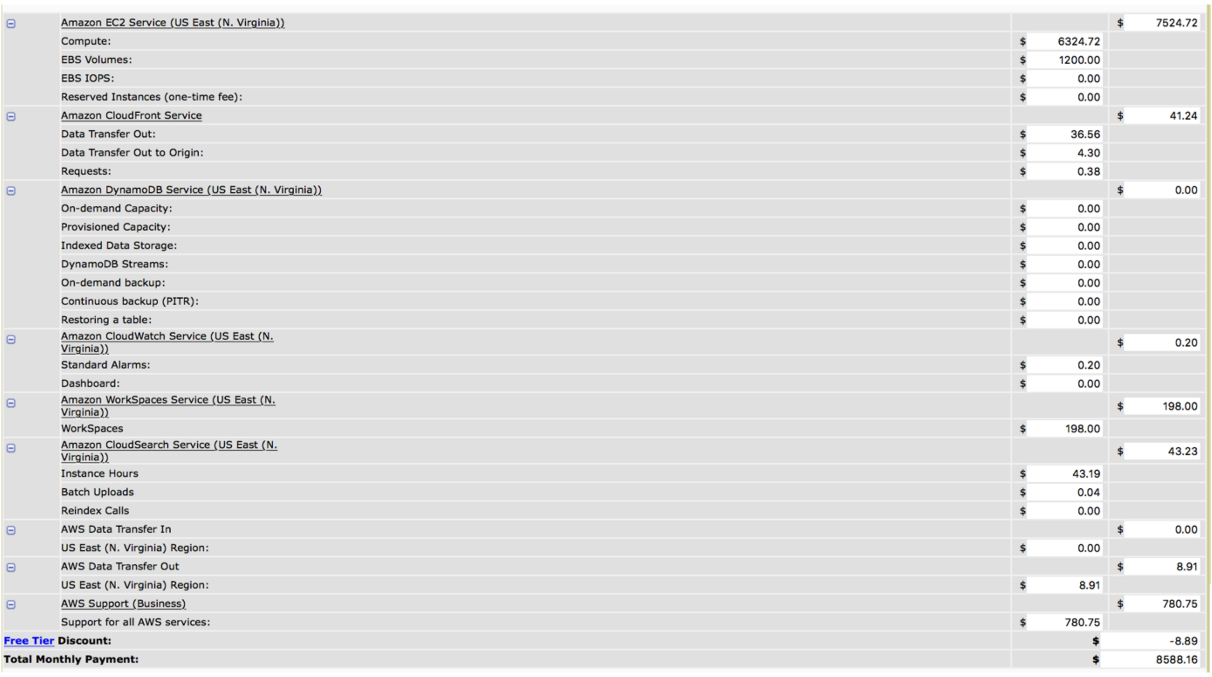
Among the business applications of AWS one can find services from Amazon WorkMail offering a secure and managed business email and calendar service to Alexa for Business promising to get more work done using this intelligent assistant. The compute area of AWS has Amazon Elastic Compute Cloud (Amazon EC2) providing secure, resizable compute activity and together with the Amazon EC2 Auto Scaling service helps you manage your EC2 instances automatically, removing or adding instances according to some defined condition. As for databases AWS offers a bunch of different options from high performance managed relational databases of Amazon Aurora to Amazon DynamoDB a managed NoSQL Database service. The developers in our team can also look forward to the developer tools on AWS for example the AWS CodeBuild, AWS CodePipeline and AWS CodeDeploy for compiling source codes, automation of pipelines and fully managed deployment services. The machine learning engineers in our company can also make use of a vast array of machine learning tools from Amazon SageMaker, a fully managed service that provides building, training and deploying of ML models to using TensorFlow on AWS for deep learning in the cloud.

Since we have to make sure that our data is only accessed by authorized personnel, we can exploit AWS Security, Identity and Compliance services like AWS Firewall Manager a security management service that allows for central configuration and management of firewall rules.

With all these tempting services we have to be careful and plan those resources wisely in order not to exceed operational costs. The IT department will have to establish a management plan for all the resources and is still responsible for configuring and managing the virtual resources let alone for training the people who will be using the tools.

In Figure 2, I describe the calculation of a few services that the company can exploit using AWS. For the compute, I have allocated 12 VMs, 6 Linux and 6 Windows, on high performance servers located in US East (N. Virginia) region. Using this service accounts for most of the costs. Amazon CloudFront Service was added to ensure seamless delivery of our content using Amazon S3. With Amazon DynamoDB we can add a high performance non-relational database service with up to 25 GB of free storage. The web service Amazon CloudWatch enables us to monitor our Amazon EC2 instances and EBS volumes in real time. Our 12 AWS resources will be monitored at 1-second intervals using 1 dashboard and 12 alarms for a negligible amount of cost because it comes with a free tier. For our IT department at Cambridge there will be six Amazon WorkSpaces, fully managed desktop computing services in the cloud, on Windows licenses included for a monthly cost of 198$. The fully-managed search service Amazon CloudSearch should help us to integrate fast and highly scalable search functionality. Altogether, the monthly bill would account for 8599.16$.

To conclude, Amazon AWS provides attractive services for our company, but we need to monitor which services we are consuming and when our demand for computing is changing, we need to act in order to keep the costs at bay.

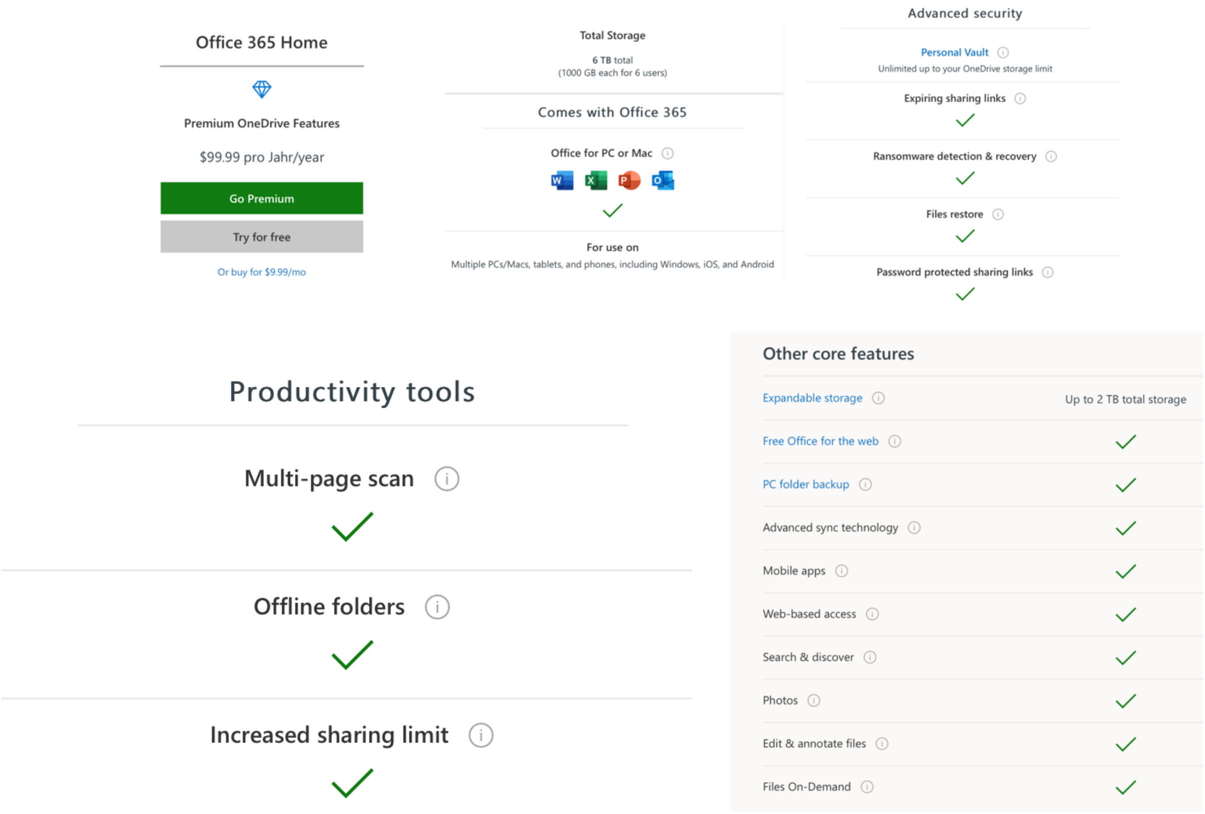


⊟	Amazon EC2 Service (US East (N. Virginia))		$ 7524.72
	Compute:	$ 6324.72	
	EBS Volumes:	$ 1200.00	
	EBS IOPS:	$ 0.00	
	Reserved Instances (one-time fee):	$ 0.00	
⊟	Amazon CloudFront Service		$ 41.24
	Data Transfer Out:	$ 36.56	
	Data Transfer Out to Origin:	$ 4.30	
	Requests:	$ 0.38	
⊟	Amazon DynamoDB Service (US East (N. Virginia))		$ 0.00
	On-demand Capacity:	$ 0.00	
	Provisioned Capacity:	$ 0.00	
	Indexed Data Storage:	$ 0.00	
	DynamoDB Streams:	$ 0.00	
	On-demand backup:	$ 0.00	
	Continuous backup (PITR):	$ 0.00	
	Restoring a table:	$ 0.00	
⊟	Amazon CloudWatch Service (US East (N. Virginia))		$ 0.20
	Standard Alarms:	$ 0.20	
	Dashboard:	$ 0.00	
⊟	Amazon WorkSpaces Service (US East (N. Virginia))		$ 198.00
	WorkSpaces	$ 198.00	
⊟	Amazon CloudSearch Service (US East (N. Virginia))		$ 43.23
	Instance Hours	$ 43.19	
	Batch Uploads	$ 0.04	
	Reindex Calls	$ 0.00	
⊟	AWS Data Transfer In		$ 0.00
	US East (N. Virginia) Region:	$ 0.00	
⊟	AWS Data Transfer Out		$ 8.91
	US East (N. Virginia) Region:	$ 8.91	
⊟	AWS Support (Business)		$ 780.75
	Support for all AWS services:	$ 780.75	
Free Tier Discount:		**$**	-8.89
Total Monthly Payment:		**$**	8588.16

Figure 2. Calculated monthly costs for selected AWS services.

Scenario 3: Online store owner

I am the owner of an online store who is selling hand painted Easter eggs. With this product I have natural changes in demand across a year with the highest sales between March and April when Easter is around. In order to deal with the great demands during that time and to cope with the quieter months, I need a flexible IT infrastructure that is able to scale to demands. For that reason, I have decided to try a Software as a Service (SaaS) cloud solution. Since I don´t have the knowledge and the time to develop my own solutions, I would like to rely on ready-made services. SaaS is basically a software solution hosted in the cloud and represents the most mature cloud service. Benefits include the on-demand delivery model which is hosted by a service provider and which can be accessed using the internet. I personally don´t care where the servers are located nor am I interested in who manages them. I just want to use specific services without getting into much of the underlying details of the technology. The model of deployment is public as I am not dealing with sensitive data and I am fine with the level of security it comes with.
Microsoft Office 365 seems to be just perfect for my needs. The main tools that are relevant for my business include the familiar Word, Excel, PowerPoint and Outlook services. With Microsoft Office 365 Home Edition I can access these tools from any of my devices at any time. For 99.99$ I get a yearly subscription of Microsoft Office 365 and I can download and install the applications immediately without having to go to the shop. The license even includes 6 TB of storage and it can accommodate up to six users. An advanced security set up such as password protected sharing links and ransomware detection & recovery can help me to keep my data save. In my business I need to take good photos to present my goods. Microsoft Office 365 Photos allows my photos to be synced with all my devices having my pictures readily available.
Altogether, Microsoft Office 365 provides the model of choice for my business.



Office 365 Home

Premium OneDrive Features

$99.99 pro Jahr/year

Go Premium

Try for free

Or buy for $9.99/mo

Total Storage

6 TB total
(1000 GB each for 6 users)

Comes with Office 365

Office for PC or Mac ✓

For use on
Multiple PCs/Macs, tablets, and phones, including Windows, iOS, and Android

Advanced security

Personal Vault
Unlimited up to your OneDrive storage limit

Expiring sharing links ✓

Ransomware detection & recovery ✓

Files restore ✓

Password protected sharing links ✓

Productivity tools

Multi-page scan ✓

Offline folders ✓

Increased sharing limit ✓

Other core features

Expandable storage	Up to 2 TB total storage
Free Office for the web	✓
PC folder backup	✓
Advanced sync technology	✓
Mobile apps	✓
Web-based access	✓
Search & discover	✓
Photos	✓
Edit & annotate files	✓
Files On-Demand	✓

Figure 3. Screenshots depicting the tools available in Microsoft Office 365 Home.