Assessment 1 Transcript

[Slide 1]

Thank you all for coming. Today I’ll be providing you with an overview of AWS or Amazon Web Services – what it is, what it can offer, an overview of its machine learning capabilities, and of course, some of its restrictions and limitations.

Before we begin, I hope you’ve had a chance to look over the glossary I sent with the presentation invite, as it’ll help with some of the AWS terminology I use.

Did you know that there are more than one million different organisations and individuals using AWS, and they encompass just about every type of industry currently in existence; including 21st Century Fox, 3M, Netflix, Automobili Lamborghini, AirBnB and NASA? (Amazon Web Services [AWS], n.d.-c)

AWS is a subsidiary of Amazon.com and offers on-demand cloud-based products and services on a pay-as-you-go basis. The service was officially launched back in 2006, and whilst there is more than one story about how exactly it came about (Butler, 2015; Miller, 2016), we do know that it arose out of a vision of a standardised and scalable company IT infrastructure that could then be marketed to others outside the company (Butler, 2015).

[Slide 2]

AWS has turned out to be a highly lucrative venture for Amazon. As of last year, AWS made up only 12% of Amazon’s total revenue, but is the parent company’s primary source of operating revenue (currently accounting for about 59% of total operating profits) (Amazon.com, 2021).

Furthermore, AWS has the largest share by far of the $130 billion global cloud-computing market at around 32%. Its two biggest rivals – Microsoft Azure and Google Cloud hold shares of approximately 20% and 9%, with smaller players including the likes of Alibaba, IBM, Salesforce and Oracle (Richter, 2021).

[Slide 3]

Whilst the AWS platform is built on four key capabilities: compute (virtual computers), storage, databases and networking, there are also services specific to media, the Internet of Things, cost management and of course machine learning and analytics. At the moment, there are over 100 services in AWS’s service offering across 25 different categories (AWS, n.d.-g).

[Slide 4]

The next key component is that of security. AWS operates a shared responsibility model, which essentially means that AWS takes care of the security of the underlying cloud infrastructure – that is, data centres, services, networking and so on – anything that Amazon manages as part of its cloud-service, but the customer takes care of the security of the services they use – such as operating systems, user access and data (AWS, 2020). A good way to think of it is that AWS manages the security **of** the cloud and the customer manages the security of **whatever they put in** the cloud (Carpenter, 2020a).

As a result, the customer gets to use a pre-established data centre and network architecture, but at the same time retain the flexibility to configure the most appropriate security controls for their own operating environment (AWS, 2020).

[Slide 5]

The final key component relates to the concepts of regions and availability zones. An AWS region is a specific geographical area of the world, within which sits collections of data centres grouped into what are known as availability zones. There are currently 25 regions, of which nine are located in the Asia-Pacific (AWS, n.d.-f; AWS, n.d.-j). According to AWS, the use of multiple data centres and availability centres increases fault-tolerance of the system and the availability of services to customers (AWS, n.d.-j).

Importantly, what this means is that a user is able to establish applications and services in whichever geographical regions best suit their needs to, for example, ensure that stored data remains in a particular country to meet government and/or industry regulations.

[Slide 6]

The next part of my presentation will be focussing on AWS’s machine learning offering – which currently includes around 20 different services. This is headed by Amazon SageMaker, a general-purpose fully-managed service that can be used to build, train and deploy machine learning models at scale with a variety of pre-built algorithms (AWS, n.d.-i). However, the remaining services have been grouped into nine different major categories or applications including Computer Vision, Artificial Intelligence, and Automated Data Extraction and Analysis. For example, Amazon Comprehend is a natural language processing service that extracts insights and relationships from text data, Amazon Fraud Detector can help identify potentially fraudulent activities, Amazon Lookout for Vision uses Computer Vision to identify missing components in products, damage to vehicles and or structures, or any other physical item where quality is important, and AWS DeepRacer, DeepLens and DeepComposer provide learning platforms for reinforcement learning, computer vision and generative AI (AWS, n.d.-i).

[Slide 7]

AWS natively supports nine different programming languages, including Ruby, JavaScript, C++ and Go (AWS, n.d.-e) through individual software development kits. However, only Python and R are natively supported by Amazon SageMaker’s notebook kernels. With respect to R, the reticulate package provides an interface into the Python SDK for R users and R’s paws library allows the user to make API calls to AWS services, similar to how the boto3 module operates in Python.

[Slide 8]

AWS supports a wide variety of machine and deep learning frameworks through the use of built-in frameworks and deep learning containers, such as scikit-learn, Apache MXNet and Spark, TensorFlow, Hugging Face (for Natural Language Processing) and Chainer, but also provides the flexibility of setting up custom-built deep learning engines or repositories for unsupported programming languages and frameworks (AWS, n.d.-b; AWS, n.d.-d).

[Slide 9]

Amazon SageMaker provides 17 different built-in algorithms that can be used for supervised and unsupervised learning tasks, dimensionality reduction, text analysis and image processing. These range from general purpose algorithms like the Linear Learning Algorithm that can be used for both classification and regression problems and K-Means clustering, to more specialised algorithms like the DeepAR Forecasting Algorithm that forecasts one-dimensional time series using recurrent neural networks (AWS, n.d.-b). It is also possible to use a script to train a customised algorithm within a supported framework, and for the ultimate flexibility and customisation, the user can create and use a new Docker container containing their own training and inference algorithms (AWS, n.d.-b).

[Slide 10]

The final part of the puzzle is the computing resources or Amazon Elastic Compute Cloud (EC2) instances that will be used in the machine learning workflow. Instances refer to the combination of CPU, memory, storage and networking, and usually come in a variety of sizes (with an associated change in price as well) (AWS, n.d.-a). Whilst there are numerous types of instances available, many of them are not suited to machine learning and AWS recommends starting with the general-purpose m5, compute-optimised c5 or accelerated p3 instance types. Furthermore, AWS notes that use of GPU instances can reduce training times significantly, which can make them more cost-effective despite the higher (and sometimes significantly higher) per-instance costs (AWS, n.d.-b).

[Slide 11]

The process of building, training and deploying a machine learning model in Amazon SageMaker is relatively simple, especially if one of the pre-built algorithms is being used. There are also a lot of commonalities with the normal machine learning workflow.

* Step 1 involves creating a jupyter notebook instance that will be used to download and process the data. As part of this, the user creates an Identity and Access Management (IAM) role that will allow Amazon SageMaker to access data held in Amazon S3.
* Step 2 involves pre-processing the dataset and then uploading the training and test sets into Amazon S3.
* Step 3 is where the model is trained. The exact form of the estimator function, including the arguments passed to it will depend on whether a pre-built algorithm is being used or a script containing a customised algorithm built outside of Amazon SageMaker. The process of tuning any relevant hyperparameters also occurs in this step. The trained model along with any other model artifacts are usually then saved back into S3.
* Step 4 is the deployment stage. The model is deployed on a server and an endpoint is created that the user can access. It is from this endpoint that additional code can be run to create predictions, both on test sets and on a real-time basis, and evaluate model performance.

(Amazon, 2021; AWS, 2021)

[Slide 12]

To summarise, some of the major advantages to using a cloud-service such as AWS for a machine learning project include:

* The ability to scale resources up and down as required depending on facts such as the type and size of datasets being analysed and the algorithms being used, and therefore only having to pay for what is used.
* Being able to take advantage of a pre-established and reliable IT infrastructure that is already configured to satisfy the requirements of customers with highly-sensitive data.
* A wide range of options depending on the level of customisation required, from pre-built algorithms to an ability to support a fully customised virtual machine (AWS, n.d.-h).
* Access to some of the latest frameworks and algorithms available, with additional functionality being added on a regular basis.

But there are also always downsides, and these include:

* Loss of control over back-end IT infrastructure.
* The possibility of service downtime for any number of different reasons.
* The risk of data breaches and information theft – arguably unlikely but still possible.
* This is not a plug-and-play platform. Education and training are recommended, particularly in the interest of avoiding unnecessary charges early on and to avoid getting overwhelmed with all the options. (Altexsoft, 2021; Larkin, 2019).

And finally, there are limitations:

* AWS has implemented service quotas for each account to help guarantee the availability of resources and to prevent the accidental provisioning of more resources than actually required. Many of these can be lifted, but multiple requests will be required if the quota needs to be lifted for more than one region (AWS, 2020), and there are some quotas that cannot be lifted under any circumstances (Sabo, 2017).
* Every minute spend on the platform will be costing money, unless a non-on demand pricing model is used, so it will generally pay to be a fast programmer.
* Not all services are available in every region. (Zou, 2020)
* There is no support for third-party packages with the built-in algorithms and only those on PyPi are supported if using script mode. (AWS, n.d.-b)

And that’s it. Questions?

.

**References**

Altexsoft. (2021, April 25). *Comparing machine learning as a service: Amazon, Microsoft Azure, Google Cloud AI, IBM Watson*. <https://bit.ly/3tGDAFY>

Amazon. (2021). *Amazon SageMaker Python SDK*. <https://bit.ly/3blq8AX>

Amazon Web Services. (2020, January 22). *Introduction to AWS security - AWS whitepaper*. <https://amzn.to/3fcaZmD>

Amazon Web Services. (2020, June 1). *How do I manage my AWS service quotas?* <https://amzn.to/3hmsovA>

Amazon Web Services. (2021). *Build, train, and deploy a machine learning model with Amazon SageMaker*. <https://amzn.to/2Re9ZX1>

Amazon Web Services. (n.d.-a). *Amazon EC2 instance types*. <https://amzn.to/2R4DCKD>

Amazon Web Services. (n.d.-b). *Amazon SageMaker: Developer guide*. <https://amzn.to/3tFfcEC>

Amazon Web Services. (n.d.-c). *AWS Customer Success*. <https://amzn.to/3o7MwD1>

Amazon Web Services. (n.d.-d). *AWS Deep Learning AMIs*. <https://amzn.to/3eDE1MQ>

Amazon Web Services. (n.d.-e). *AWS developer center*. <https://aws.amazon.com/developer/>

Amazon Web Services. (n.d.-f). *AWS global infrastructure*. <https://amzn.to/3biB4iU>

Amazon Web Services. (n.d.-g). *AWS management console - AWS services*. <https://amzn.to/33BZaRq>

Amazon Web Services. (n.d.-h). *Benefits at a glance*. <https://amzn.to/3hsaqro>

Amazon Web Services. (n.d.-i). *Machine learning on AWS*. <https://amzn.to/3eDEfDG>

Amazon Web Services. (n.d.-j). *Regions and availability zones: Region maps and edge networks*. <https://amzn.to/2ReGm8b>

Amazon.com. (2021, February 3). *United States Securities and Exchange Commission Form 10-K*. <https://bit.ly/33yobwT>

Butler, B. (2015, March 2). *The myth about how Amazon's web service started just won't die*. <https://bit.ly/3vRQdzs>

Carpenter, T. (2020, May 27). *AWS certified solutions architect - associate (SAA-C02): 1 cloud services overview | Security and compliance*, [Video]. LinkedIn Learning. <https://bit.ly/2Q9L46w>

Larkin, A. (2019, August 7). Disadvantages of cloud computing. *Cloud Adoption*. <https://bit.ly/2RNcbVk>

Miller, R. (2016, July 3). *How AWS came to be*. <https://tcrn.ch/3eDdXRV>

Richter, F. (2021, February 4). *Amazon leads $130-billion cloud market*. <https://bit.ly/3eDOcBh>

Sabo, I. (2017, January 31). 5 AWS limitations every CEO needs to be aware of. *Cloud Adoption*. <https://cloudacademy.com/blog/5-aws-limitations-to-be-aware-of/>

Zou, J. (2020, July 6). *A gentle introduction to AWS SageMaker - ML & AI on the cloud*. <https://bit.ly/3faA5SQ>