Pearson

Introduction to Amazon Web Services

Tomasz Matysiak &
Mateusz Otmianowski
Advanced Computing & Data Science Lab

WiMLDS, 16.01.2019

Image by: Tang Yau Hoong

Github repo with notebooks:

https://github.com/MateuszOtmianowski/wimlds_aws

Agenda

- AWS in general: idea, history
- AWS services (theory + exercises)
  - S3
  - EC2
  - Athena

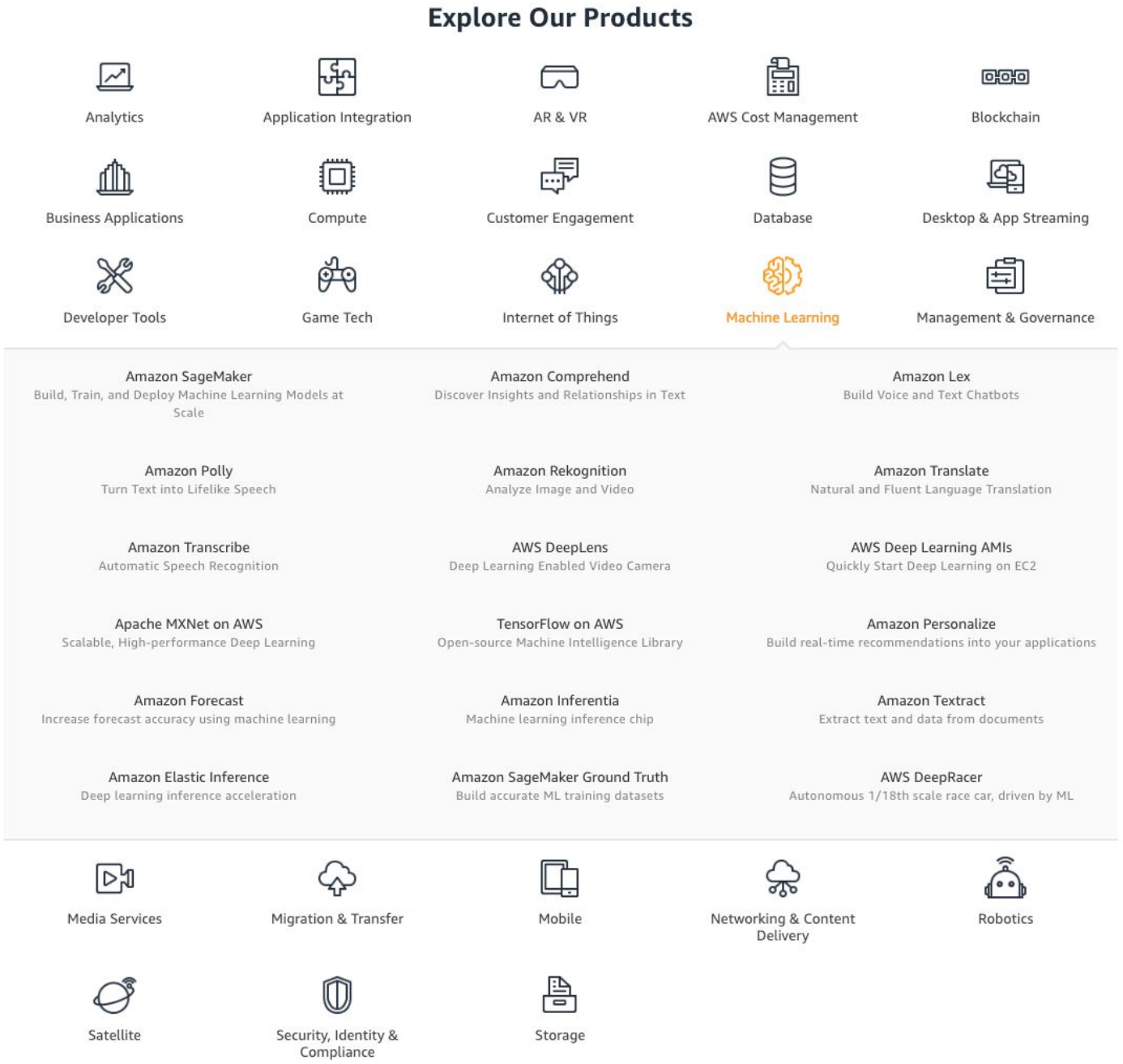
Amazon Web Services

What is AWS ?

“Amazon Web Services (AWS) is a subsidiary of Amazon that provides on-demand cloud computing platforms to individuals, companies and governments, on a paid subscription basis.”

AWS services

Image from: https://aws.amazon.com/products/



S3
Simple Storage Service

What is S3?

- Amazon Simple Storage Service (Amazon S3) is
  - an object storage service that offers scalability, data availability, security, and performance
  - dedicated for customers of all sizes
  - one of two services started by Amazon in 2006

Advantages

- **Secure**: data encryption, access management tools, monitoring access
- **Cost effectiveness**: cheaper than Hadoop according to AWS engineers, technologies like S3 Storage Class
- **Durability:** 99.999999999%
- **Scalable**
- **APIs**: Easy to integrate with lots of applications

Disadvantages

- **Significant costs of transfer** between various regions
- **High region dependance**
- **May be expensive in some cases:** IO costs of S3 archive (Glacier)

S3 Storage Classes

Image source: https://aws.amazon.com/s3/storage-classes/

Performance across the S3 Storage Classes

	S3 Standard	S3 Intelligent-Tiering*	S3 Standard-IA	S3 One Zone-IA†	S3 Glacier	S3 Glacier Deep Archive**
Designed for durability	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)
Designed for availability	99.99%	99.9%	99.9%	99.5%	N/A	N/A
Availability SLA	99.9%	99%	99%	99%	N/A	N/A
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved
First byte latency	milliseconds	millseconds	milliseconds	milliseconds	select minutes or hours	select hours
Storage type	Object	Object	Object	Object	Object	Object
Lifecycle transitions	Yes	Yes	Yes	Yes	Yes	Yes

Features

- Data stored as objects in so-called **buckets**
- Flat structure, no directories, no folders
- **Prefixes**
- **Naming convention**
    - bucket: ‘my-bucket’
    - prefix: ‘project_name/data/redshift/2018/01/accepted/’
    - object name: ‘test.txt’
  - Full path
    - my-bucket/project_name/data/redshift/2018/01/accepted/test.txt
- **No copying, no object renaming**
- **Version-control**
- **Pricing policy,** pay-per use

No renaming on S3, just copy-pasting

- Example
    - We have an object
        - my-bucket/project_name/data/redshift/2018/01/**accepted**/test.txt
    - I want to ‘move’ file to:
        - my-bucket/project_name/data/redshift/2018/01/**to_delete**/test.txt

A trap: file have to be copy-pasted. Problematic for huge files!

EC2

What is EC2?

- Amazon Elastic Compute Cloud (Amazon EC2) is
  - a web service that provides secure, resizable compute capacity in the cloud
  - designed to make large-scale cloud computing easier for developers
  - one of two services started by Amazon in 2006

Advantages

- **Easy to set up**: standard, preconfigured machines
- **API access**: easy to use
- **Resizable**: quick scaling up and down, minutes to multiply your computing resources, auto-scaling options
- **Security**: IAM accounts, permissions policy, security groups
- **Virtual Private Cloud**
- **Pricing policy**: pay-per use
- **Regions:** choose the place you operate in
- **Free tier**

Disadvantages

- **In-build limits** of every service (per region)
- **Expensive technical support:** ranging from 29 $ for single developer up to dozens of thousands $ per big enterprise
- **Limited customization**: no option to add just RAM or CPU to your instance - need to run new bigger instance
- **Relatively expensive** in ‘run-on-demand’
- Difficult to **estimate** all costs

Features

- **Storage volumes:** storage for EC2 instances, independent
- **AMIs:** pre-configured templates
  a. AMI lifecycle

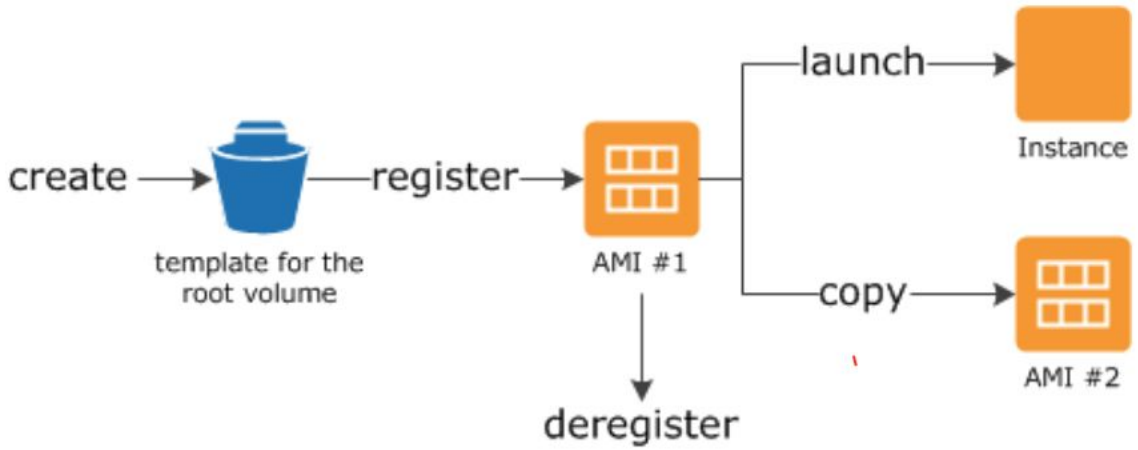


Image from: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AMIs.html

Features

- **Various types of machines:** including highly-efficient GPU instances
- **Regions and availability zones:** physical place of storage
- **Amazon Cloudwatch:** statistics
- **Operating systems:**

Image from: https://aws.amazon.com/ec2/features/

Pricing

- **Per second billing:**
  - Minimum 1 minute, after that billing per every second
- **Running the instance:**
  - On-demand
    - Depends on region and on instance type
    - Price range
      - 0,006$ - 16$ per hour
      - 4,32$ - 11.520 $ per month
      - 52,56 $ - 140.160 $ per year
  - Spot-prices
    - Up to 90% cheaper compared to On-demand
    - May be interrupted during the operation
  - Reserved instances
    - Up to 75% cheaper compared to On-demand
    - Dedicated for at least 1 one year of usage

Instance types

- **General purpose** (A1, T3, T2, M5, M5a, M4)
  - web servers, microservices, dev environments, low-latency interactive applications, small and medium databases, virtual desktops, cluster computing
- **Compute optimised** (C5, C5n, C4)
  - high performed web servers, scientific modelling, batch processing, distributed analytics, multiplayer gaming
- **Memory optimised** (R5, R5a, R4, X1e, X1)
  - memory intensive applications such as high performance databases, distributed web scale in-memory caches, mid-size in-memory databases, real time big data analytics
- **Accelerated computing** (P3, P2, G3, F1)
  - machine/deep learning, high performance computing, speech recognition, 3D visualizations, graphics-intensive remote workstation, 3D rendering, application streaming
- **Storage optimised** (H1, I3, D2)
  - MapReduce-based workloads, distributed file systems such as HDFS and MapR-FS, network file systems, log or data processing applications such as Apache Kafka, and big data workload clusters

ALWAYS LEARNING