

University of  
South Wales  
Prifysgol  
De Cymru

MS4S21 Big Data Engineering  
and Applications  
  
Week 4  
  
Moizzah Asif  
moizzah.asif@southwales.ac.uk  
J418

Moizzah Asif - Big Data Engineering and Applications
© University of South Wales

1

---

---

---

---

---

---

---

---

University of  
South Wales  
Prifysgol  
De Cymru

## Recap

### Overview of

- The need for big data technologies
- Popular big data storage models
- Popular data models
- Virtual machine creation
- Linux (ubuntu) terminal commands
- Envisaging a big data Hadoop/HDFS cluster
- Setting up machines for a HDFS cluster
- Envisaging a mapreduce job on a HDFS cluster
- Looking at a standalone local map function
- Implementing a mapreduce job on a HDFS cluster

2
Moizzah Asif - Big Data Engineering and Applications
© University of South Wales

2

---

---

---

---

---

---

---

---

University of  
South Wales  
Prifysgol  
De Cymru

## This Week

### Big Data Cloud services?

### Case study AWS



3
Moizzah Asif - Big Data Engineering and Applications
© University of South Wales

3

---

---

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## This Week

### Big Data Cloud services



#### Case Study: Big Data through Amazon AWS cloud

- The whats?
- The hows?
- The whens?

4 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

4

---

---

---

---

---

---

---

---


University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

Broad Analytics Usage In The AWS Cloud


Being in the cloud gives us the scalability of adding application and database servers as we need them.

Keith Mitchell  
Programmer, reddit.com



Using AWS, we have increased our reliability by an order of magnitude.

Yuri Izraelvsky  
VP of Cloud and Platform Engineering, Netflix



5 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

5

---

---

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

AWS let us build solutions for an environment that moves so rapidly, you can't plan for it.

Michael Staley  
Chief Integration and Innovation Officer, OFA



**MS4S21 : Big Data Moizzah**

10 MAR 2021

### US Re-elections - Win Win for AWS and Obama

the organisation that drove Barack Obama's 2012 campaign re-election as president of the United States.

OFA needed an inexpensive, highly available, scalable system to serve millions of users.

Using the AWS cloud helped save the OFA campaign tens of million dollars in hardware costs.

Obama campaign used AWS to coordinate thousands of volunteers saved millions in hardware costs and fundraise online.

Obama for America (OFA) is the organisation that drove Barack Obama's 2012 campaign re-election as president of the United States.

OFA needed an inexpensive, highly available, scalable system to serve millions of users.

Using the AWS cloud helped save the OFA campaign tens of million dollars in hardware costs.

Obama campaign used AWS to coordinate thousands of volunteers saved millions in hardware costs and

6 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

6

---

---

---

---

---

---

---

---

University of  
South Wales  
Prifysgol  
De Cymru

## Big Data Cloud Services - AWS

### AWS Services For Big Data Workloads

Sources of Truth	Real time	High Performance Databases	Analysis Platforms
 Amazon S3 Amazon Glue Amazon Redshift	 Amazon Kinesis	 Amazon DynamoDB Amazon Aurora	 Amazon EMR

7
Moizzah Asif - Big Data Engineering and Applications
© University of South Wales

---

---

---

---

---

---

---

---

---

---


7

University of  
South Wales  
Prifysgol  
De Cymru

## Big Data Cloud Services - AWS

### Take a step back ...

Where does the big data workload run on a cloud???



- Elastic Server Capacity
- Instance Choice
  - CPU
  - Memory
  - Storage
- Deployment Options
- OS: Amazon Machine Images (AMIs)
- Applications

8
Moizzah Asif - Big Data Engineering and Applications
© University of South Wales

---

---

---

---

---

---

---

---

---

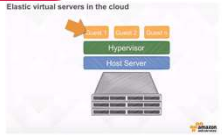
---

8

University of  
South Wales  
Prifysgol  
De Cymru

## Big Data Cloud Services - AWS

### EC2 – Compute service



- You have control of your instances
- Log on as root (Linux) / Administrator (Windows)
- Install the software you need
- Start/Stop and control via console or APIs
- Make the configuration changes you like
- Create an AMI (Amazon Machine Image)

- Choose the instance type that suits you
- Change the instance type when you want to
- Attach as much or as little storage as you need
- Choose your operating system
- Choose a pre-configured image (AMI)

9
Moizzah Asif - Big Data Engineering and Applications
© University of South Wales

---

---

---

---

---

---

---

---

---

---

9

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### EC2 – Compute service

#### Amazon EC2 Instance Families

General Purpose:	M1, M3, T2
Compute Optimized:	C1, CC2, C3, C4
Memory Optimized:	M2, CR1, R3
Dense Storage:	HS1, D2
I/O Optimized:	HI1, I2
GPU:	CG1, G2
Micro:	T1, T2

Instance generation  
**c4.large**  
Instance family      Instance size

10      Moizzah Asif - Big Data Engineering and Applications      © University of South Wales

10

---

---

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### EC2 – Compute service

#### Choose your instance types

Try different configurations to find your optimal architecture.

General	CPU	Memory	Disk/I/O
m1 family m3 family	c3 family cc1.4xlarge cc2.8xlarge	m2 family r3 family	d2 family i2 family
Batch process	Machine learning	Spark and interactive	Large HDFS

11      Moizzah Asif - Big Data Engineering and Applications      © University of South Wales

11

---

---

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### EC2 – Compute service

#### Storage types (*Block storage?*)

- Locally attached or "instance storage"
- Amazon EBS General Purpose (SSD) volumes
- Amazon EBS Provisioned IOPS (SSD) volumes
- Amazon EBS Magnetic volumes
- Amazon S3/Amazon Glacier

12      Moizzah Asif - Big Data Engineering and Applications      © University of South Wales

12

---

---

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### EC2 – Compute service

#### Cost - options

On-Demand	Reserved	Spot	Dedicated
Pay for compute capacity by the hour with no long-term commitments	Make an EC2 usage commitment & receive a significant discount	Bid for unused capacity, charged at a Spot Price which fluctuates based on supply & demand	Launch instances within Amazon VPC that run on hardware dedicated to a single customer
For spiky workloads, or to define needs	For committed utilization	For time-insensitive or transient workloads	For highly sensitive or compliance related workloads

13 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

13

---

---

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### Cloud – Storage

Elastic Block Store Amazon EBS	Simple Storage Service Amazon S3	Amazon Glacier
1GB to 16TB Volumes up to 20,000 IOPS per volume with EBS PIOPS	Highly Scalable Object Store Objects from 1 byte to 5TB 99.9999999% durability	Long term archive storage. Extremely low cost per GB 99.9999999% durability
Very Fast Block devices to attach to EC2 Instances	Fast API Accessible Object Storage	3-5 hour access latency Intended for write once, read never use-cases

14 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

14

---

---

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### EBS (Elastic Block Store)

EBS volumes are created in an in a specific Availability Zone

attached to an instance in that same Availability Zone

provides the following volume types: General Purpose SSD, Provisioned IOPS SSD, Throughput Optimized HDD, and Cold HDD

Performance metrics, such as:  
*bandwidth, throughput, latency, and average queue length,*  
 are available through the AWS Management Console and Amazon CloudWatch

15 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

15

---

---

---

---

---

---

---

---

16

---

---

---

---

---

---

17

---

---

---

---

---

---

18

---

---

---

---

---

---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### Useful weblinks

EC2:

- [Introduction to Amazon EC2 - Elastic Cloud Server & Hosting with AWS – YouTube](#)
- [Amazon EC2](#)

EBS:

- [Amazon Elastic Block Store \(EBS\) - Amazon Web Services](#)
- [Amazon Elastic Block Store \(EBS\) Overview – YouTube](#)
- [Amazon EBS Resources – Amazon Web Services](#)
- [Amazon EBS volume types - Amazon Elastic Compute Cloud](#)

S3:

- [What is Amazon S3? - Amazon Simple Storage Service](#)
- [Introduction to Amazon Simple Storage Service \(S3\) - Cloud Storage on AWS - YouTube](#)
- [Cloud Object Storage | Store & Retrieve Data Anywhere | Amazon Simple Storage Service \(S3\)](#)
- [Amazon S3 objects overview - Amazon Simple Storage Service](#)
- [Getting started with Amazon S3 - Amazon Simple Storage Service](#)
- [Creating object key names - Amazon Simple Storage Service](#)

19 Moizazah Asif - Big Data Engineering and Applications © University of South Wales

19

---

---

---

---

---

---

---


---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### NoSQL Databases

#### Dynamo DB



Fully managed NoSQL database

Keep you away from worrying about:

1. Scalability
2. Server provision
3. Cluster scaling
4. Down time
5. Having to think about in RDBMS ways

20 Moizazah Asif - Big Data Engineering and Applications © University of South Wales

20

---

---

---

---

---

---

---


---

University of South Wales  
Prifysgol De Cymru

## Big Data Cloud Services - AWS

### NoSQL Databases

#### Dynamo DB



A fast and flexible NoSQL database service  
Consistent, single-digit millisecond latency at any scale  
A fully managed cloud database  
Supports both document and key-value store models  
Flexible data model and reliable performance

[aws.amazon.com/dynamodb/](https://aws.amazon.com/dynamodb/)

21 Moizazah Asif - Big Data Engineering and Applications © University of South Wales

21

---

---

---

---

---

---

---

---

## NoSQL Databases

## Dynamo DB



### Create – Tables, items

### Add – Tables, items

### Delete – Tables, items

### Query – Tables, items

22

Moizzah Asif - Big Data Engineering and Applications

© University of South Wales

---

---

---

---

---

---

University of  
South Wales  
Prifysgol  
De Cymru

## NoSQL Databases

## Dynamo DB

## High availability and durability in DynamoDB



23

Moizzah Asif - Big Data Engineering and Applications

© University of South Wales

---

---

---

---

---

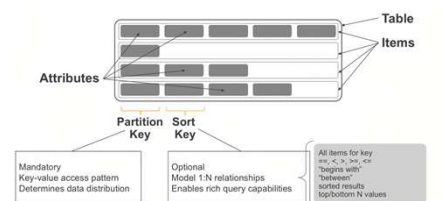
---

University of  
South Wales  
Prifysgol  
De Cymru

## NoSQL Databases

## Dynamo DB – Dissecting the tables

Table



34

Moizzah Asif - Big Data Engineering and Applications

© University of South Wales

---

---

---

---

---

---



## JSON primer

### Format

stands for JavaScript Object Notation

"self-describing" and easy to understand

language independent

JSON files can be used to store data, retrieve it through any programming language.

Packages, libraries available in

**Python:** json package <https://docs.python.org/3/library/json.html>

**R:** rjson, jsonlite etc <https://cran.r-project.org/web/packages/rjson/index.html>  
<https://cran.r-project.org/web/packages/jsonlite/index.html>

25

Moizzah Asif - Big Data Engineering and  
Applications

© University of South Wales

25

## JSON primer

### Syntax

Syntax is in name/value pairs

*"name": "Moizzah"*

Data is separated by commas

*"name": "Moizzah", "position": "Lecturer"*

JSON object is placed inside curly braces

*{"name": "Moizzah", "position": "Lecturer"}*

**Note:** The name/keys are strings, so have to be placed inside double quotes

26

Moizzah Asif - Big Data Engineering and  
Applications

© University of South Wales

26

## JSON primer

### Data types

In **JSON**, values must be one of the following data types:

- a string
- a number
- an object (JSON object)
- an array
- a boolean
- null

Python	JSON
dict	object
List, tuple	array
str	string
int, long, float	number
True	true
False	false
None	null

27

Moizzah Asif - Big Data Engineering and  
Applications

© University of South Wales

27

## JSON primer

### Examples

- a string : `{ "name": "Moizzah" }`
- a number : `{ "office": 418 }`
- an object (JSON object) : `{  
 "employee":  
 { "name": "Moizzah", "office": 418, "position": "Lecturer" }  
}`
- an array: `{  
 "employees": [ "Filippo", "Moizzah", "Penny" ]  
}`
- a Boolean: `{ "research_active": true }`
- Null : `{ "PhD_students": null }`