Teaching Big Data Analytics with AWS

Lessons Learned

Javier A. Espinosa Oviedo

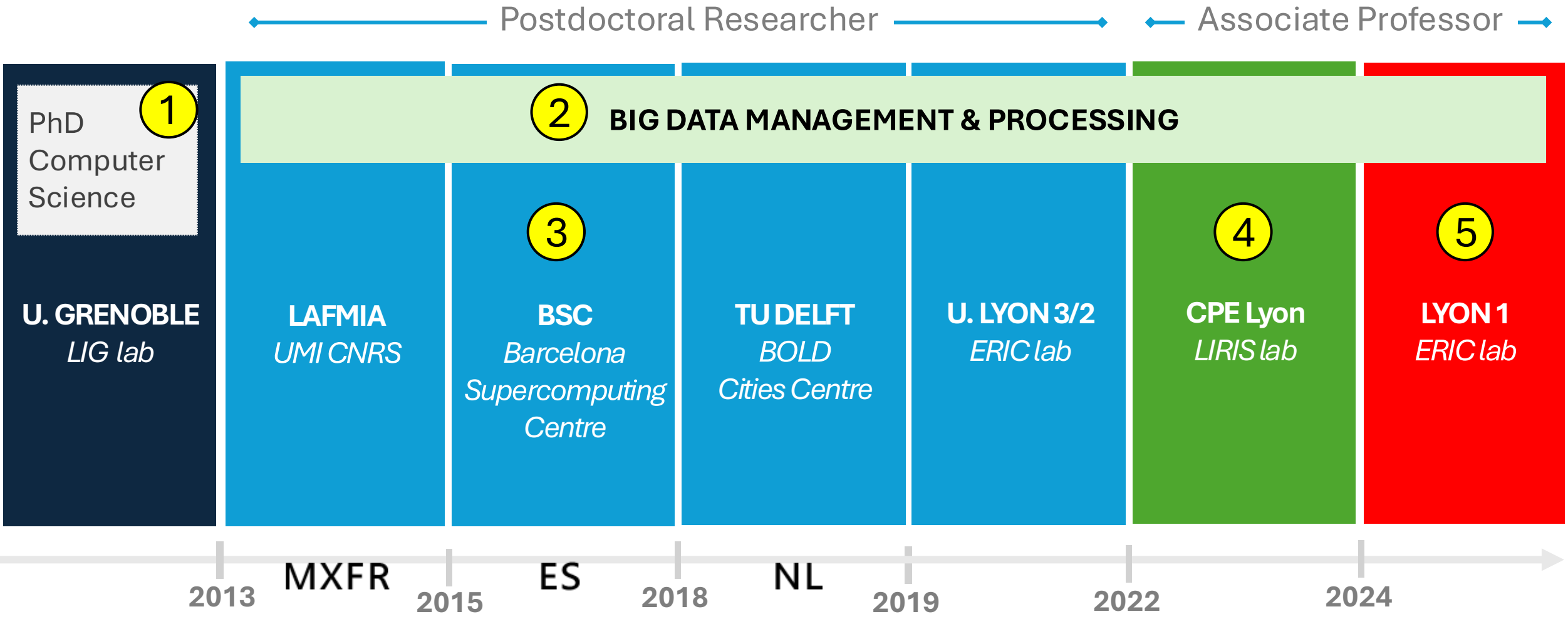
Maître de Conferences
Département d'Informatique

javier.espinosa@univ-lyon1.fr
www.espinosa-oviedo.com

June 4th, 2026

LaTêteDansLesNuages@UGA

About Me


Postdoctoral Researcher
Associate Professor
PhD Computer Science
1
2
BIG DATA MANAGEMENT & PROCESSING
3
4
5
U. GRENOBLE
LIG lab
LAFMIA
UMI CNRS
BSC
Barcelona Supercomputing Centre
TU DELFT
BOLD Cities Centre
U. LYON 3/2
ERIC lab
CPE Lyon
LIRIS lab
LYON 1
ERIC lab
2013
MXFR
2015
ES
2018
NL
2019
2022
2024

Teaching Activities

1 **Courses**

- Big Data Analytics
- Cloud Computing

2 **Programs**

- Apprenticeship track
- Research track

3 **Tools**

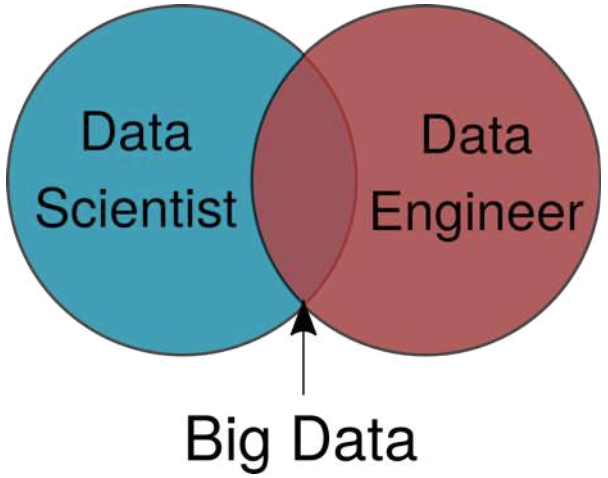
aws academy

Google Cloud

databricks University Alliance

Education

Target profiles

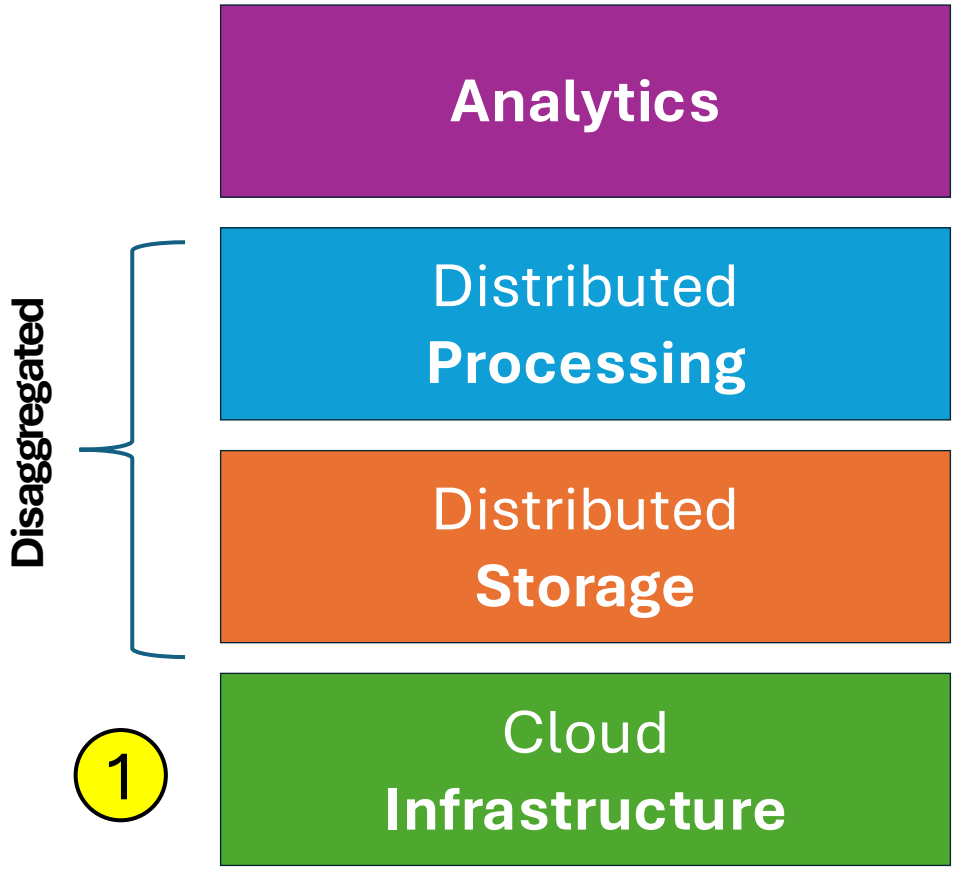


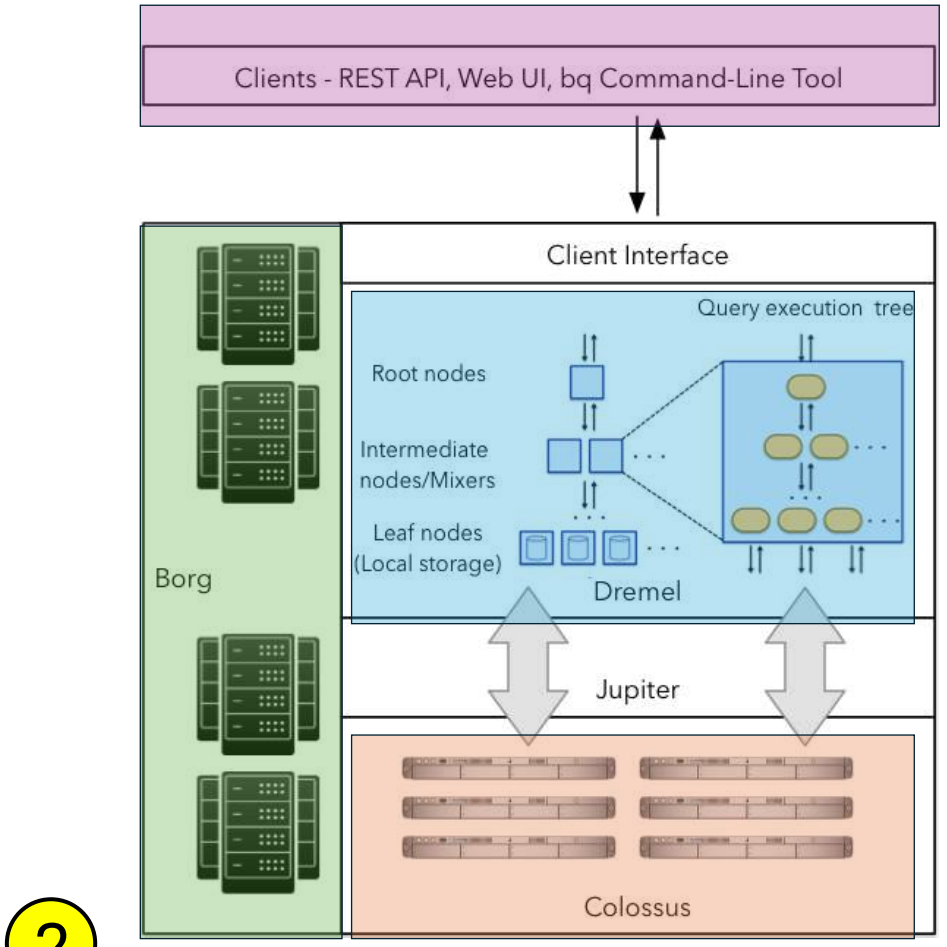
4

Objective

Provide students with a comprehensive understanding of the **Modern Data Stack**

Modern Data Stack





BigQuery Architecture
(Google data analytics platform)

Cloud Infrastructure

Analytics

Processing

Storage

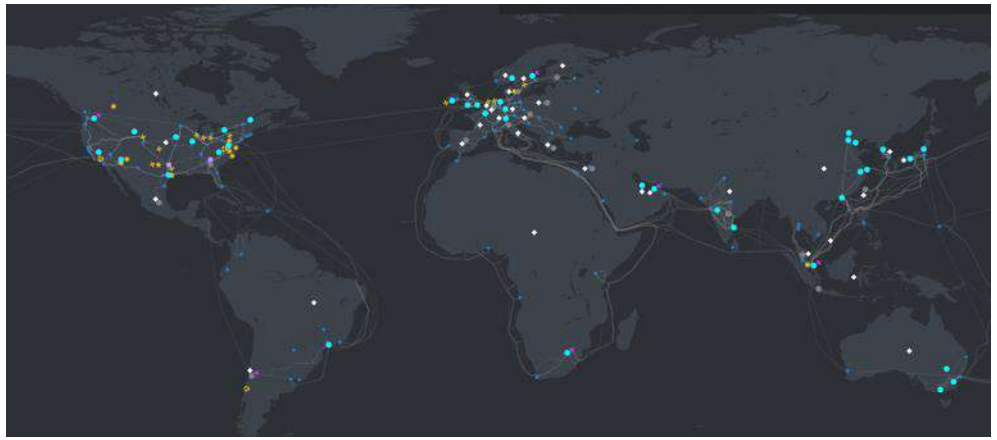
Infrastructure

1 Learning objective

- Raise awareness of the hyperscale infrastructure that powers the Modern Data Stack

2 Key concepts

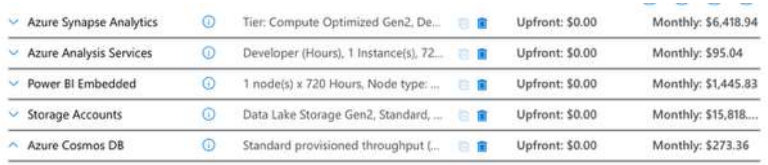
- Cloud service model (*aaS)
- Networking (VPC)
- Pricing model (pay-as-you-go)
- Global infrastructure
  - Regions
    - Availability Zones
      - Data Centers



Cloud Infrastructure: Activities

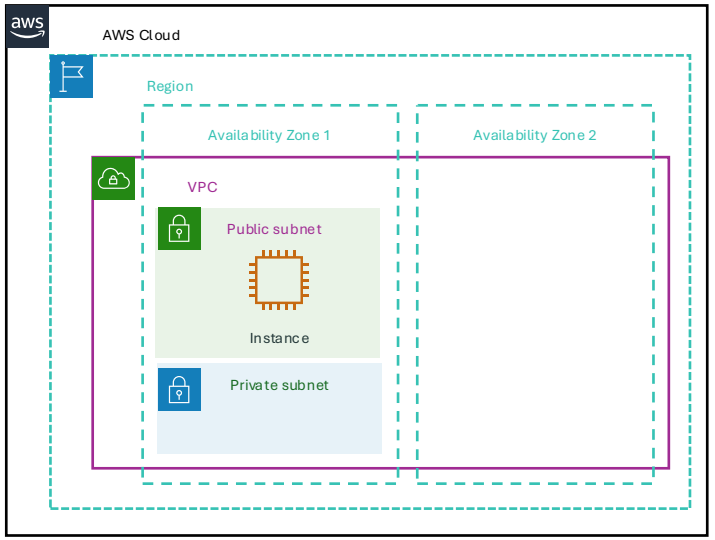
1 Cost Estimation

- Estimate the cost of a data pipeline using the AWS **Pricing Calculator**



2 Infrastructure Deployment

- Deploy an AWS **EC2 instance** within a custom VPC



Cloud Infrastructure: Activities

- Visit to the CNRS IN2P3* Computing Center

Computing Room

Mass Storage Room

Computing Museum

** Institut National de Physique Nucléaire et de Physique des Particules*

Distributed Storage

Analytics
Processing
Storage
Infrastructure

1 Learning objectives

- Understand the effect of data location and storage choices on latency, redundancy, and cost

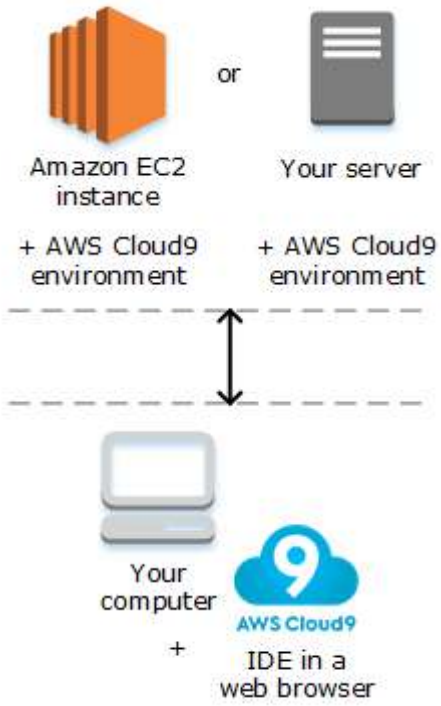
2 Key concepts

- Block Storage vs Object Storage
- Storage classes (hot, warm, cold tiers)
- Service Level Agreements
  - Availability **99.9%**
  - Durability **99.999999999%**

Distributed Storage: Activities

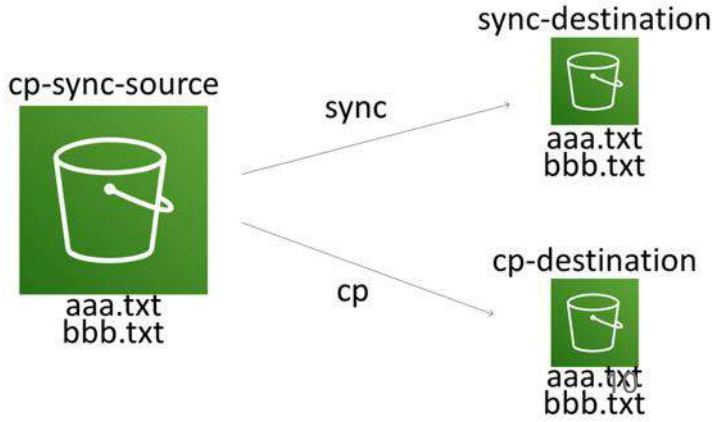
1 HDFS Fault Tolerance

- Create a HDFS cluster in AWS **Cloud9** to study the HDFS fault tolerance mechanism based on block replication



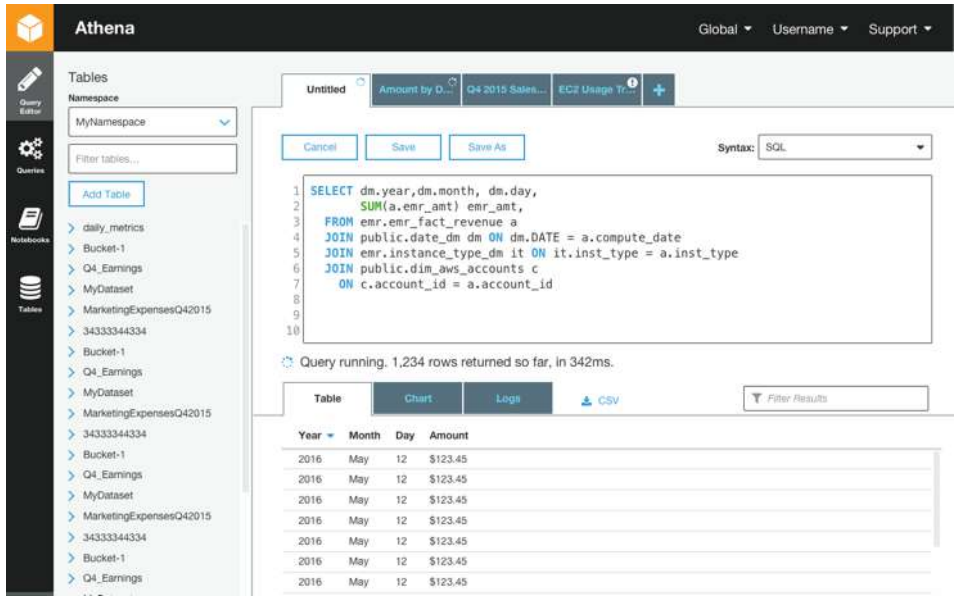
2 Amazon S3 Exploration

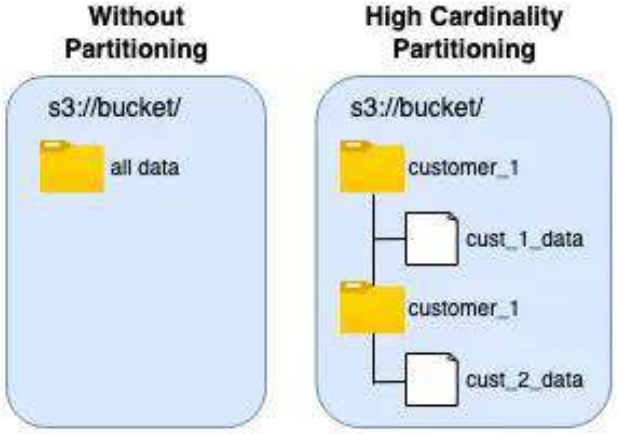
- Understand **Amazon S3** buckets and objects, and experiment with cross region replication and storage classes (S3 Standard, S3 Glacier)



Distributed Storage: Activities

- Demo: Federated queries on Amazon S3 using **Athena**
  - Pricing is based on data scanned (GB scanned per query)





Distributed Processing

Analytics
Processing
Storage
Infrastructure

1. Learning objective
   - Understand how large-scale datasets are processed in a distributed cluster environment

2. Key concepts
   - Map → Reduce data processing
   - Directed Acyclic Graph data processing
   - RDDs
   - Spark execution model & architecture
     - Jobs > Stages > Tasks
     - Driver > Executors > Slot/Core

Distributed Processing: Activities

1 Understanding MapReduce

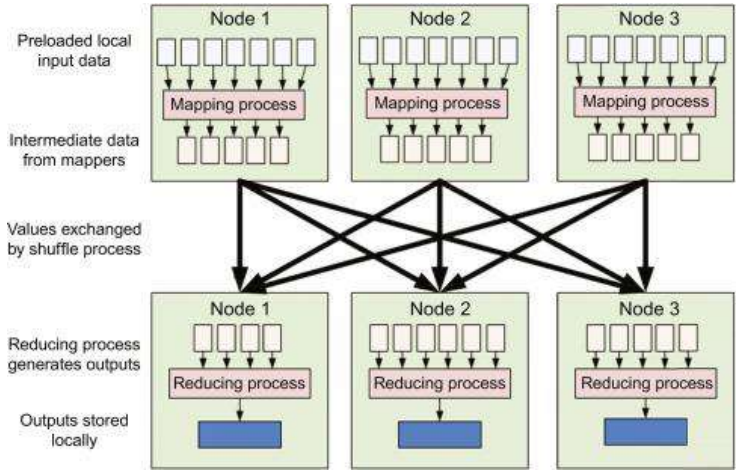
- MapReduce patterns by hand and with a simulated cluster

We Will Rock You
Song by Queen (1977)

1. Buddy, you're a boy, make a big noise
2. Playing in the street, gonna be a big man someday
3. You got mud on your face, you big disgrace
4. Kicking your can all over the place, singin

5. We will, we will rock you
6. We will, we will rock you

Post its



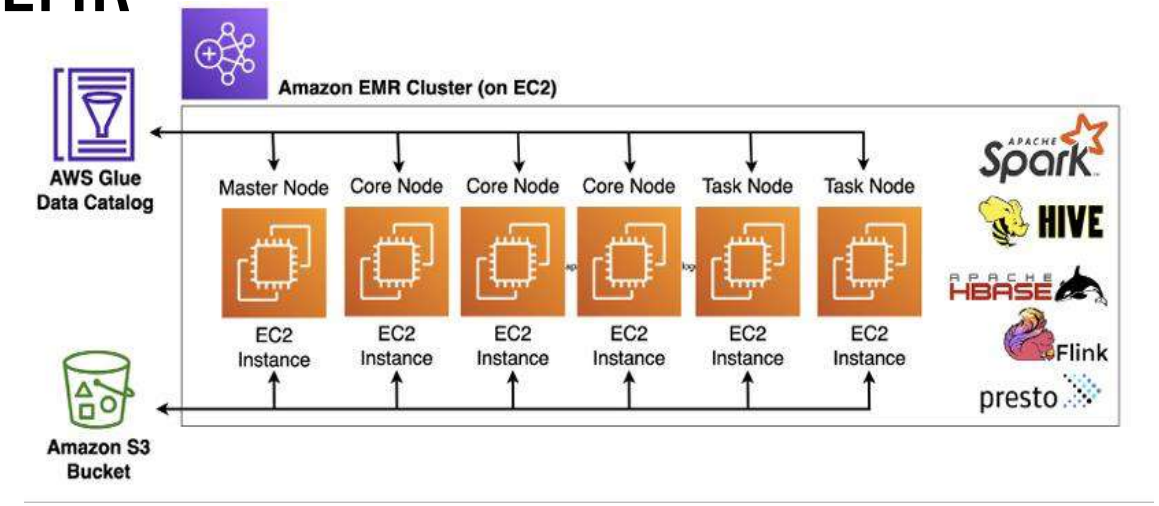
Distributed Processing: Activities

2 **Building a Spark Cluster**

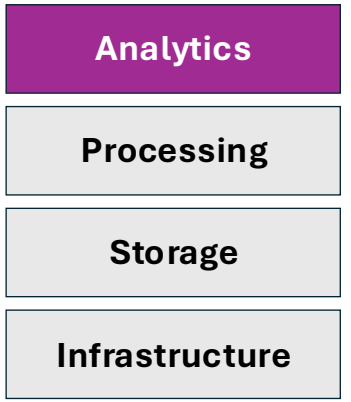
- Creating a custom spark cluster with **EC2 instances**

3 Demo: **Large-Scale Processing with AWS EMR**

- Implement a Word Count application using AWS EMR and Amazon S3



Big Data Analytics



1. Learning objective
   - Understand how modern analytics platforms such as Databricks work, and how they use distributed computing for large-scale data analysis

2. Key concepts
   - Data Frames
   - Spark SQL => Query optimizer
   - Graph / stream / semi-structured processing

Big Data Analytics: Datathon

www.espinosa-oviedo.com/big-data-analytics

Data Type	File List	#Files	Total Size Compressed (TiB)
Segments	segment.paths.gz	100	
WARC	warc.paths.gz	100000	86.48
WAT	wat.paths.gz	100000	14.27

Big Data Analytics: Datathon

web crawler

2

Collection Analysis

- Extract ALL urls
- Extract top level-domains

Text Analysis

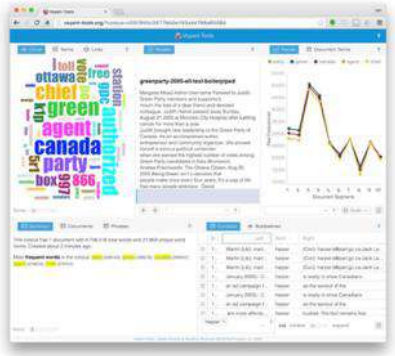
- Extract text in X language
- Extract text without boilerplate

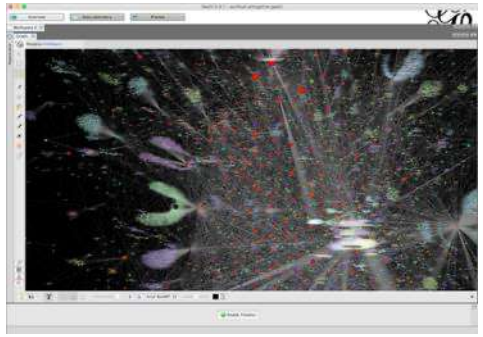
Link Analysis

- Extract link structure
- Export to Gephi

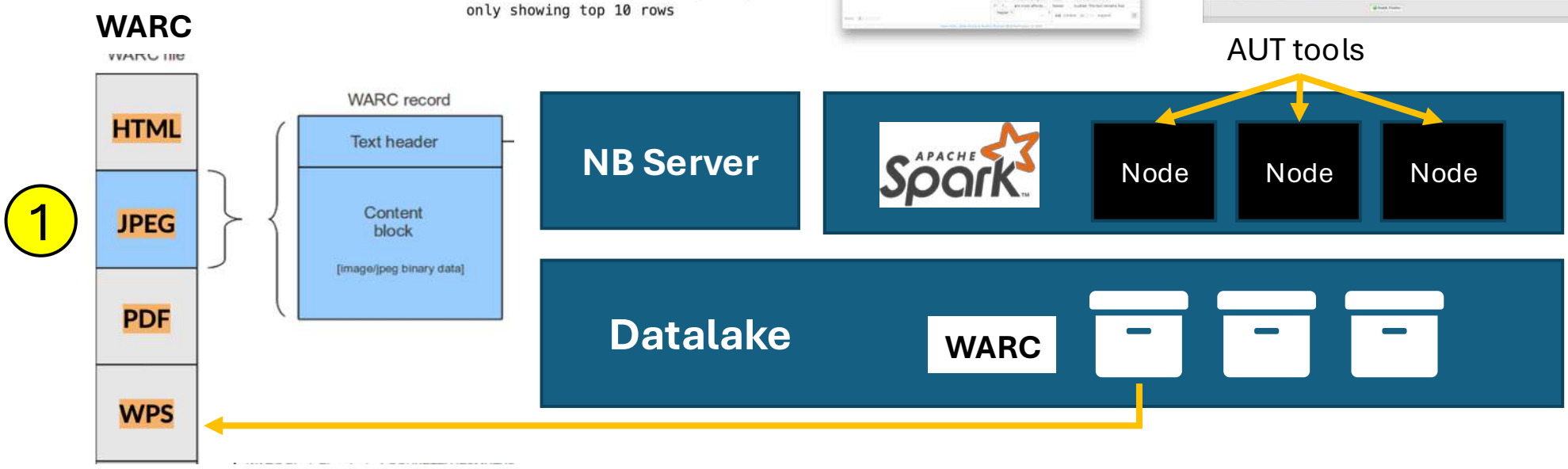
domain	count
www.equalvoice.ca	4274
www.liberal.ca	1968
www.policyalternatives.ca	588
greenparty.ca	535
www.fairvote.ca	442
www.ndp.ca	416
www.davidsuzuki.org	348
www.canadiancrc.com	88
communist-party.ca	39
www.ccsd.ca	22

only showing top 10 rows

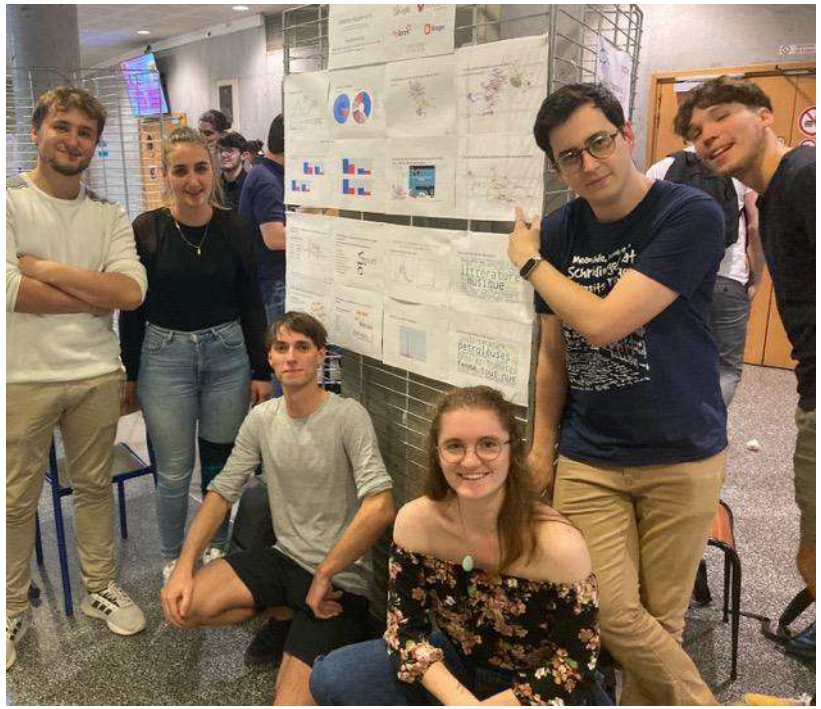


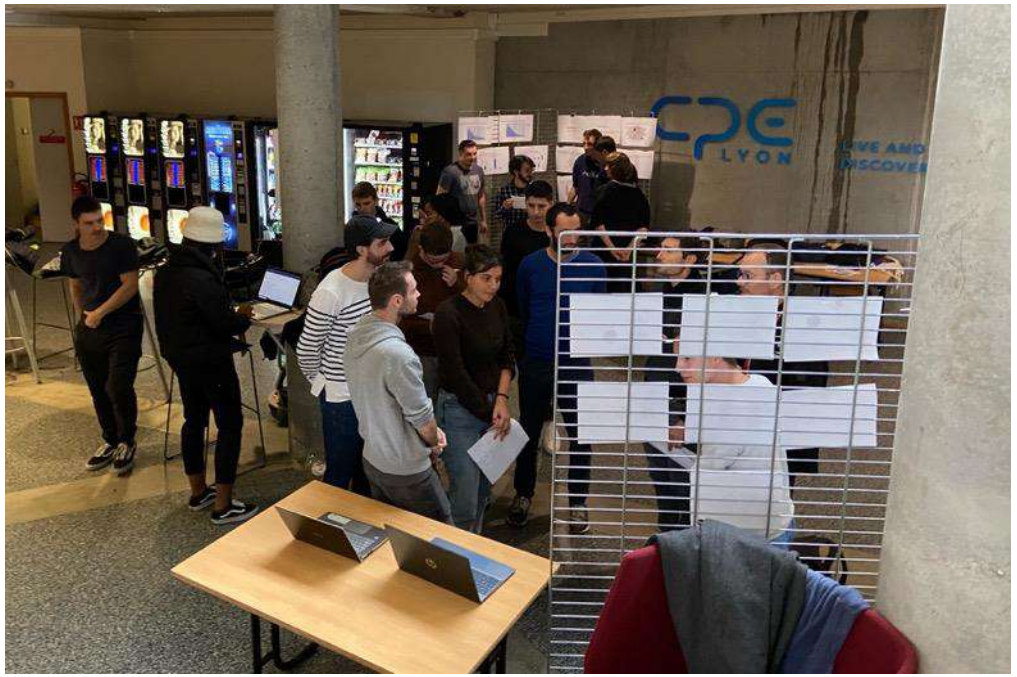


AUT tools



Big Data Analytics: DemoFest

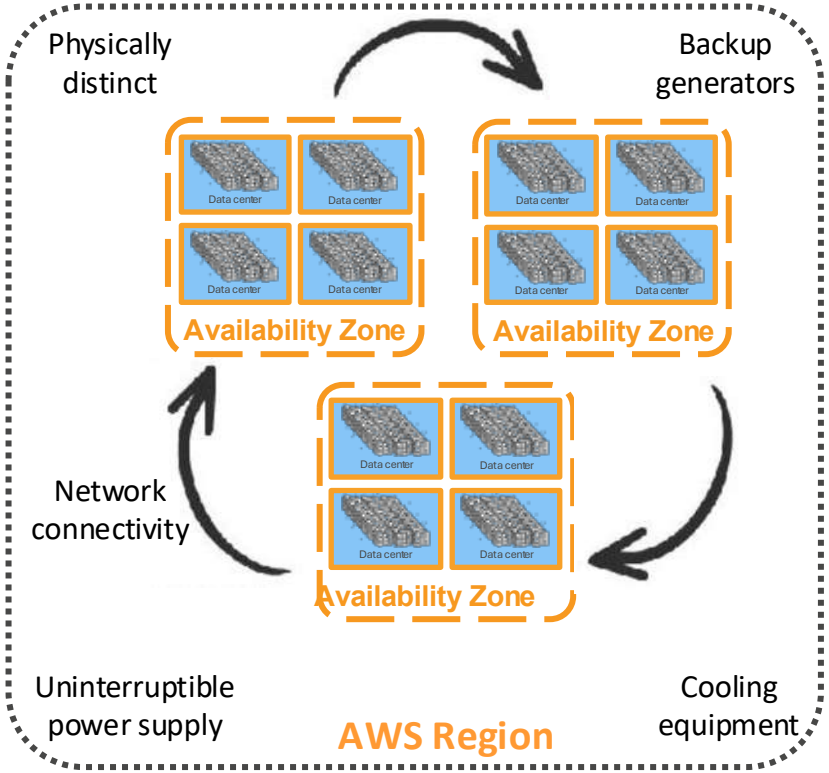


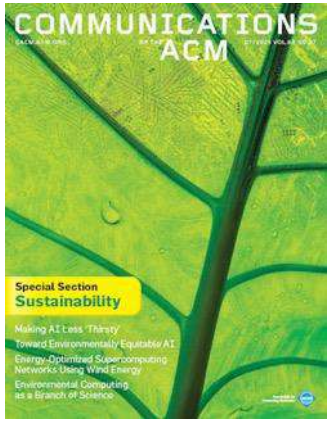


Lessons Learned

Public clouds offer a complementary perspective

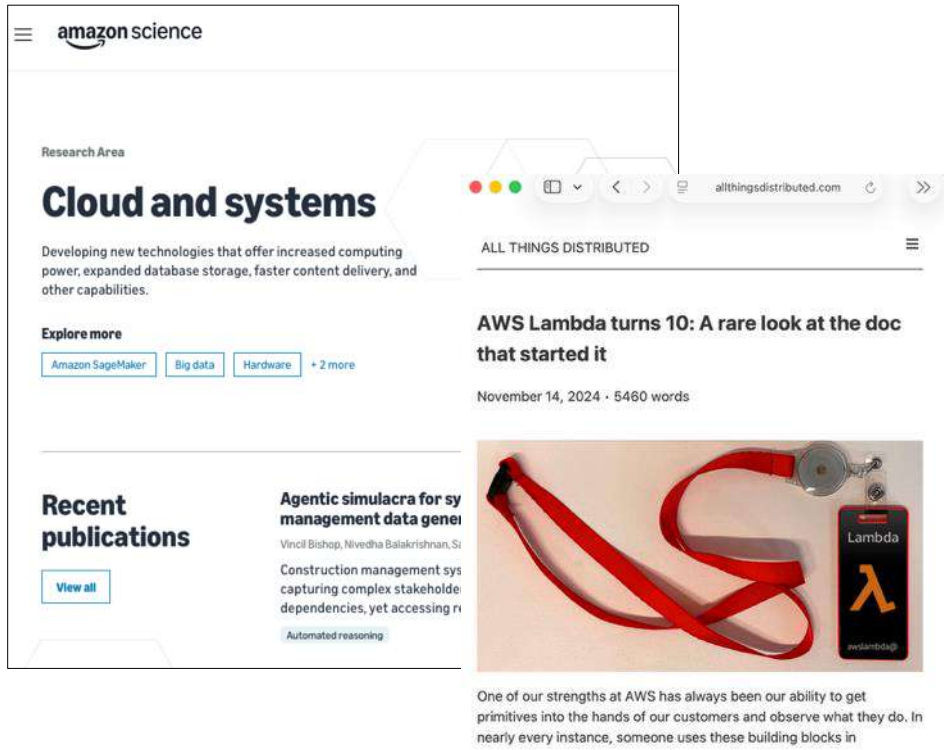
1. Data processing at scale requires elastic infrastructure
2. **Data locality** impacts performance and cost
3. High availability and fault tolerance are built into modern cloud architectures
4. Environmental footprint





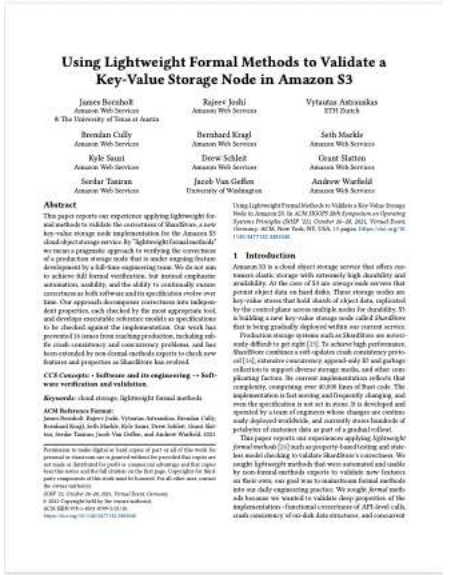
Help to connect Research & Practice

- Scientific papers and blog articles provide insights into popular cloud services internal



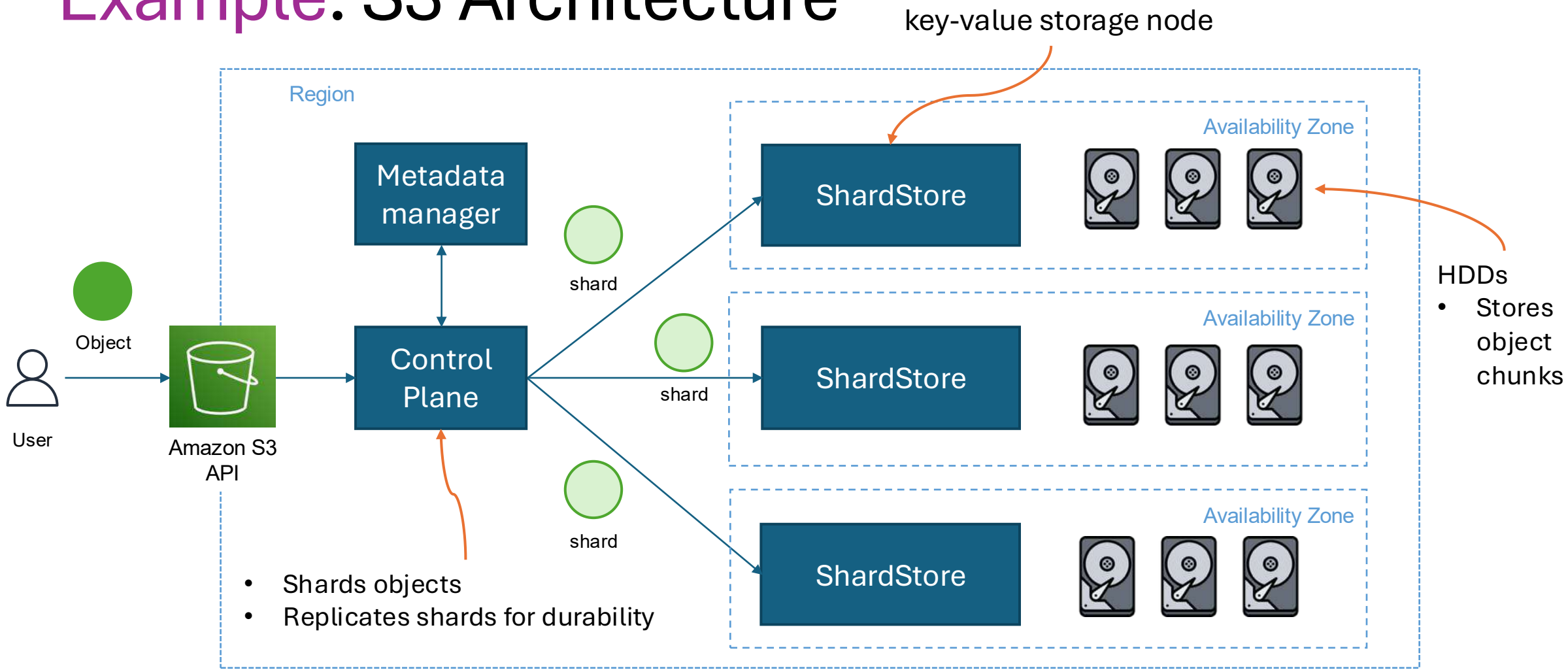

The Story of AWS Glue

AWS Glue
(it is all about spark)


Using Lightweight Formal Methods to Validate a Key-Value Storage Node in Amazon S3

AWS S3
(key-value sharding stores)

Example: S3 Architecture*



*Using Lightweight Formal Methods to Validate a Key-Value Storage Node in Amazon S3. SOSP'21, 2021

Practical exercises are “simplified”

1. Students only require a web browser
2. Students can experiment with administrative privileges
3. Extensive documentation, tutorials, and service simulators are available

4. Infrastructure is accessible 24/7 from anywhere
   - If AWS goes down, **half the Internet starts panicking**!!

Students feedback

1. Experience with cloud platforms provides a competitive advantage on the job market
2. Certification vouchers are a plus+ 

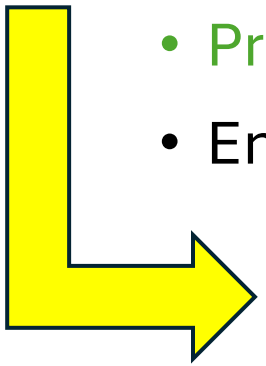
Why AWS for teaching ?

	Credits	Certification Benefit	Provided Material	Ecosystem
AWS	**50$** (prof. enroll students on demand)	4 exam vouchers	• Slides • Ready to use projects • MCQs • Multilingual	AWS build
GCP	**50$** (prof. request coupons in advance)	50% discount	• Slides	Skills boost + 5000 credits
Azure	**100$** (student redeem his/her own credits yearly)	Varies by program	-	Microsoft Learn (free)

Conclusions

1. Teaching with AWS works well, but keep it simple
   - Focus on concepts, avoid replicating production environments
   - Design exercises that can be completed within a course session

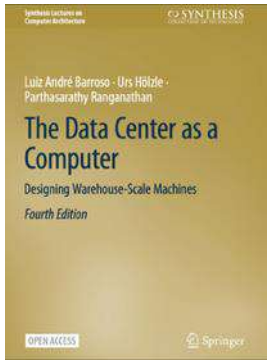
2. Academic **programs evolve** continuously
   - Prefer *free-tier* or *always-free* services
   - Ensure exercises remain viable even if educational programs change

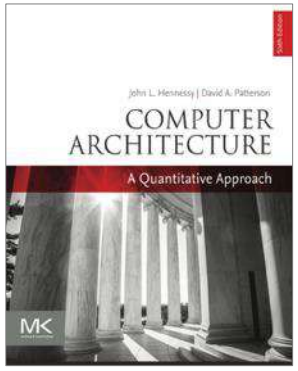


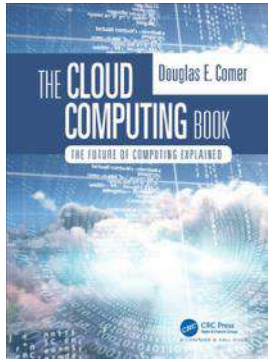
- GCP temporarily suspended educational credits outside the US in 2025
- GitHub Education reduced its Codespaces offering in July 2026

Recommended Resources

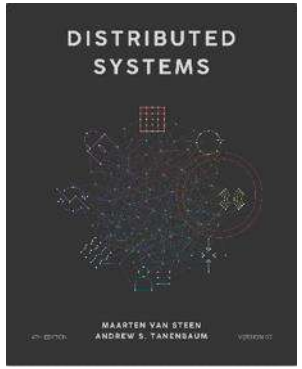
Free

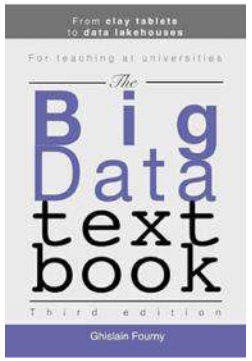




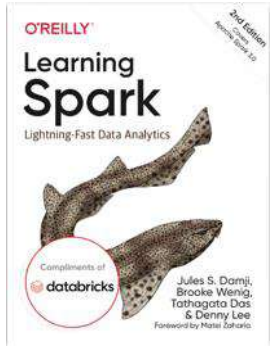


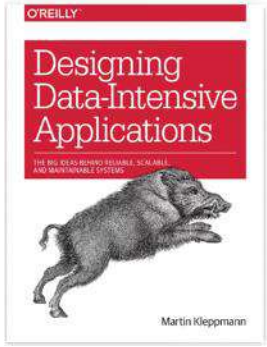
Free

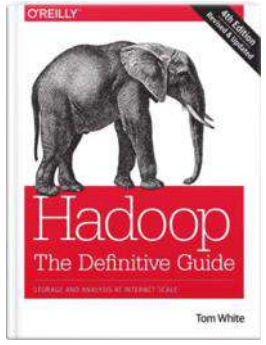


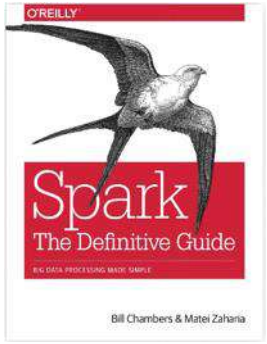


Free





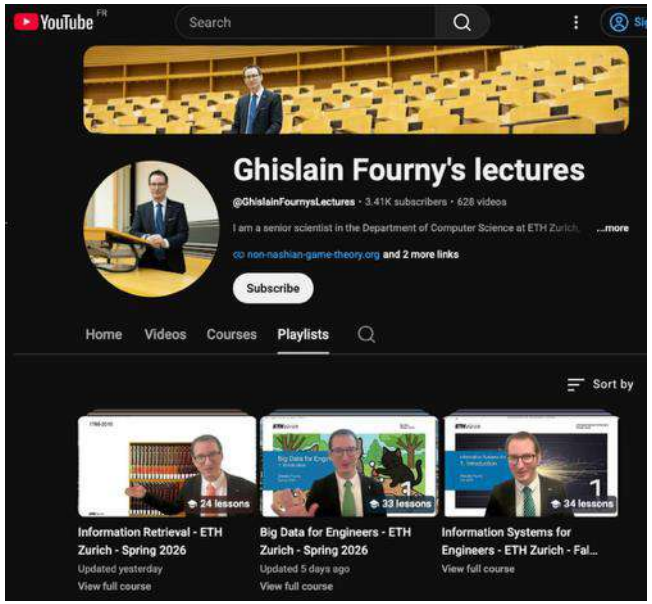




Free (formerly)

Lectures

- @GhislainFournysLectures



Merci
Questions ?