

Teaching Big Data Analytics with AWS

Lessons Learned

Javier A. Espinosa Oviedo

Maître de Conférences
Département d'Informatique

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

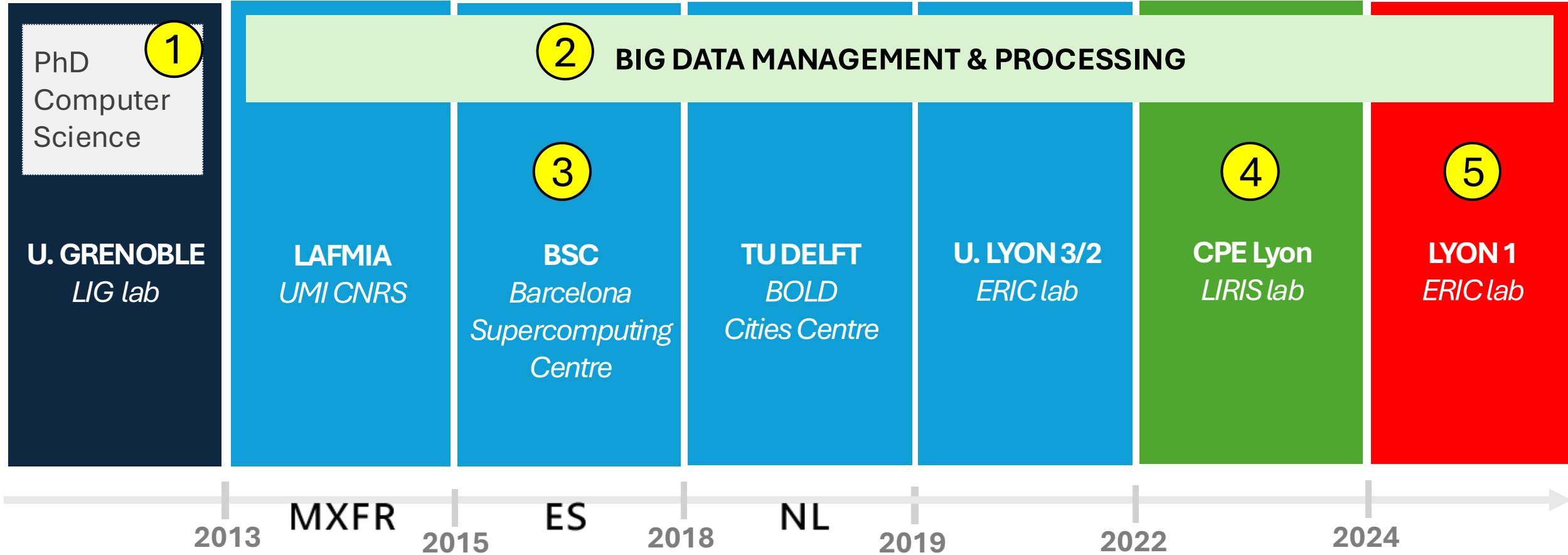
June 4th, 2026

LaTêteDansLesNuages@UGA

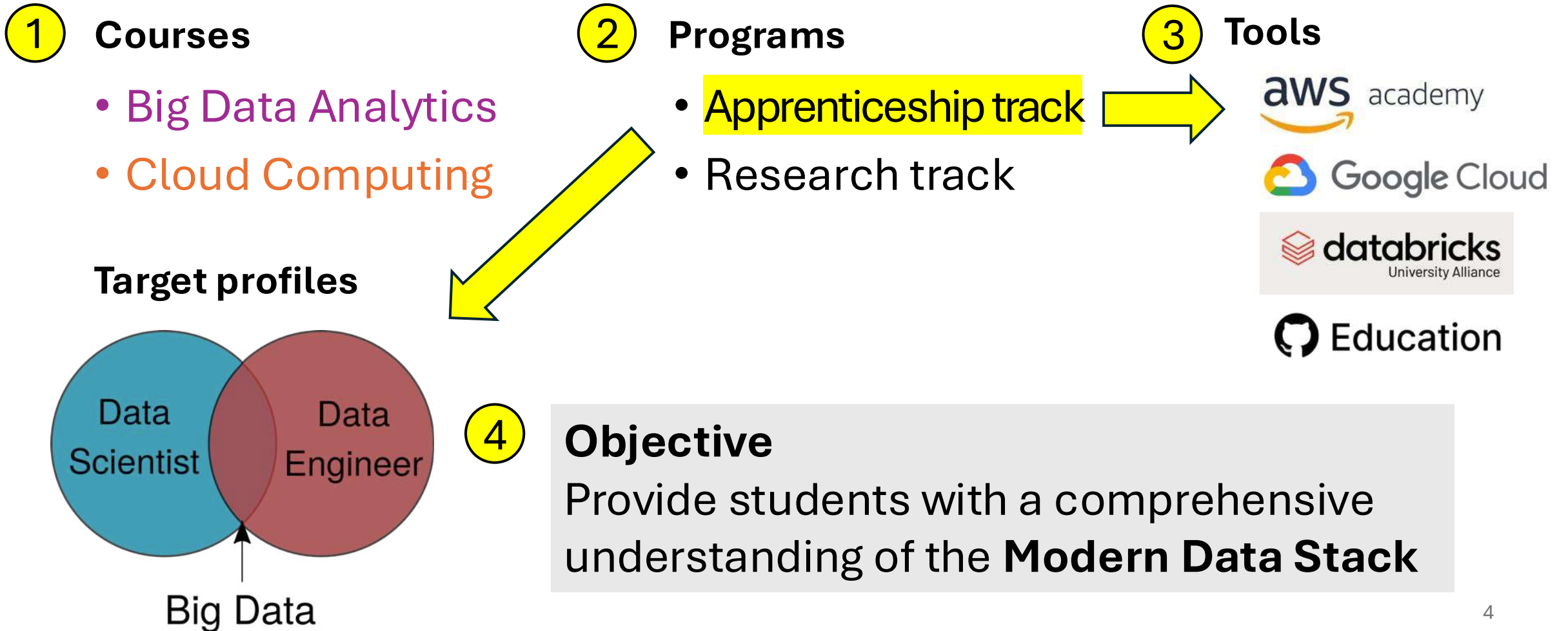


About Me

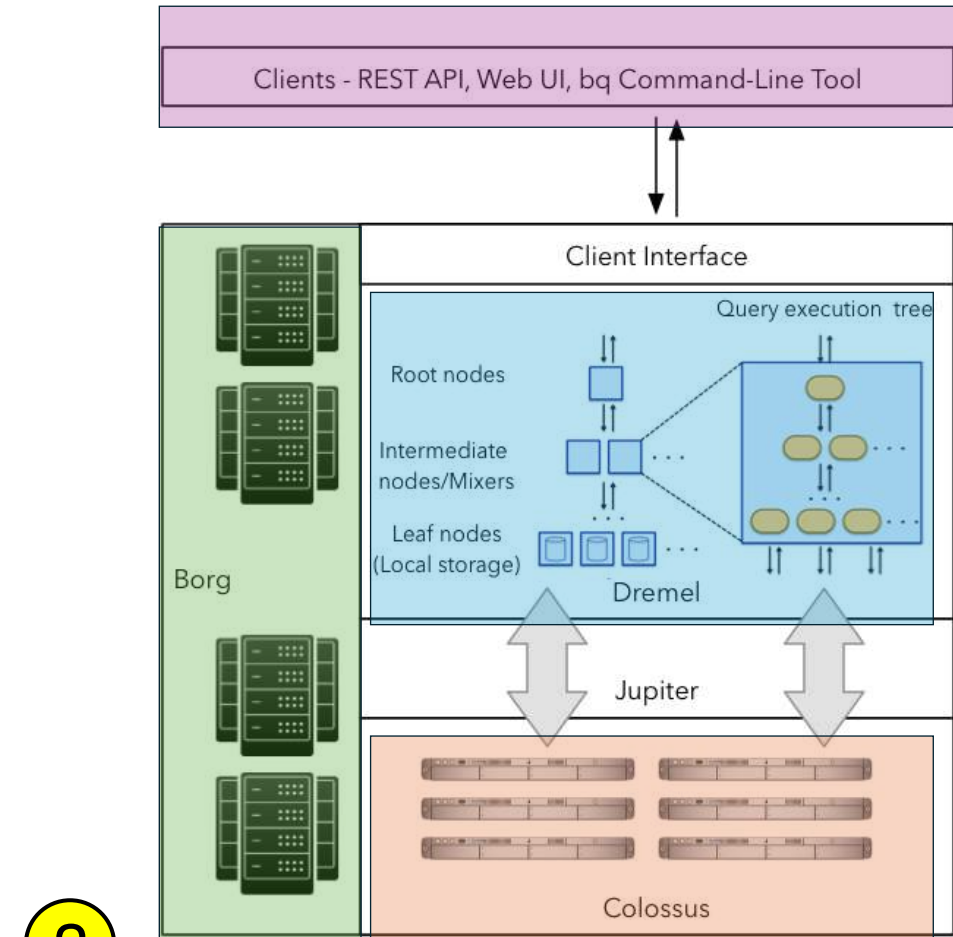
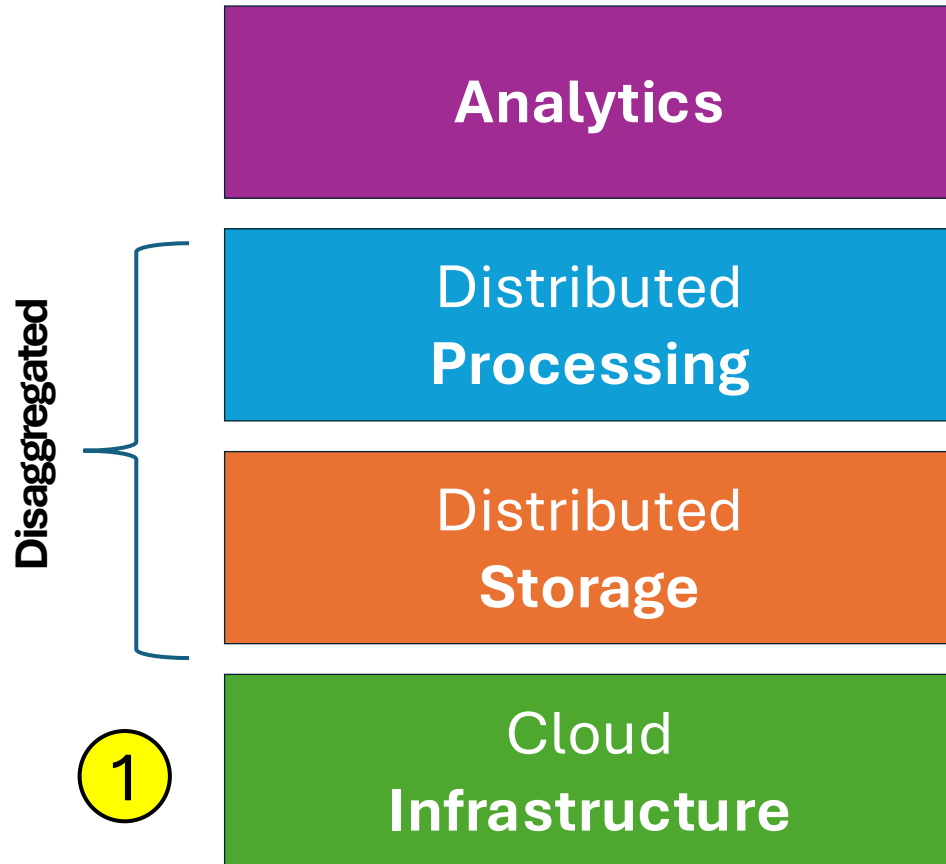
← Postdoctoral Researcher → ← Associate Professor →



Teaching Activities

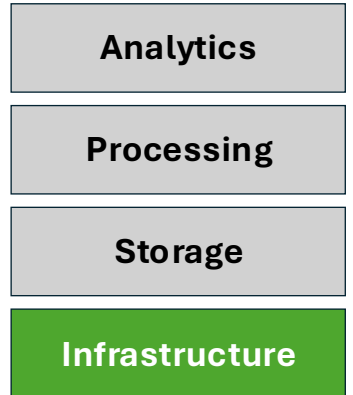


Modern Data Stack

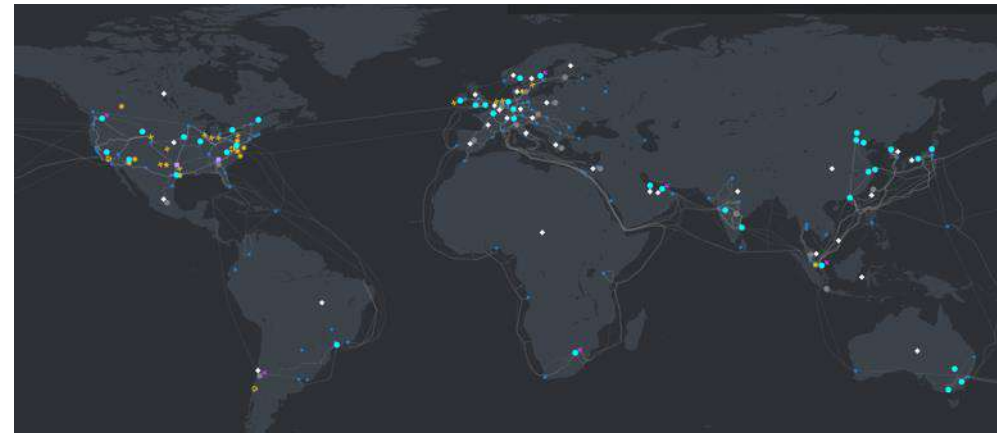


BigQuery Architecture
(Google data analytics platform)

Cloud 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**

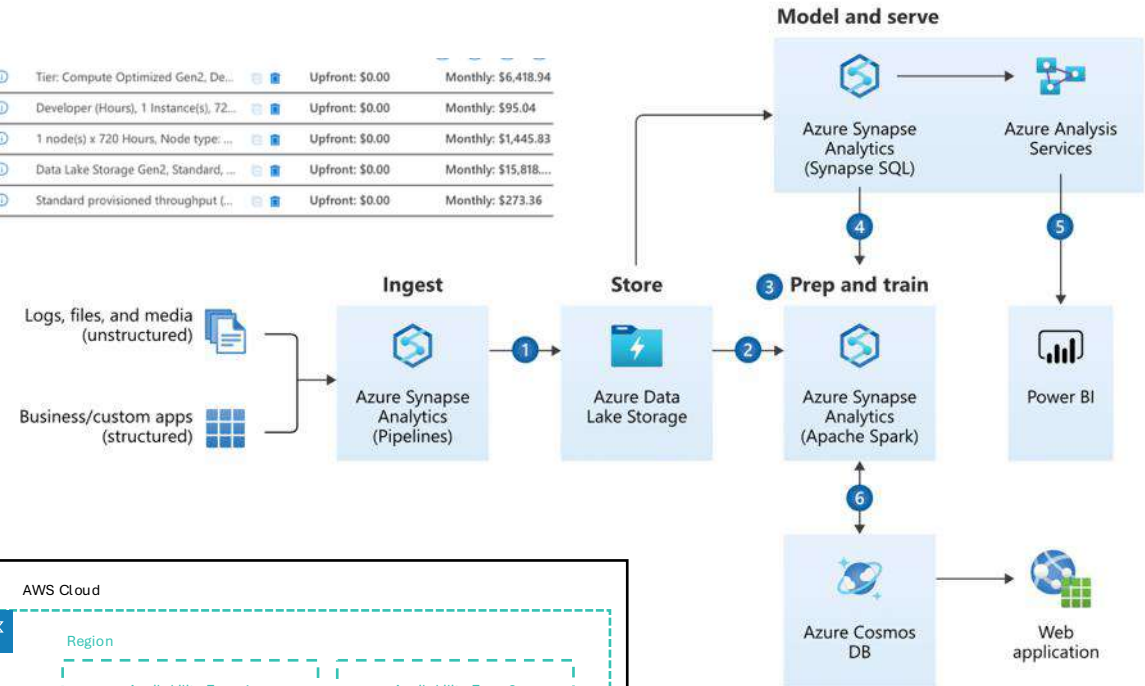


Cloud Infrastructure: Activities

1 Cost Estimation

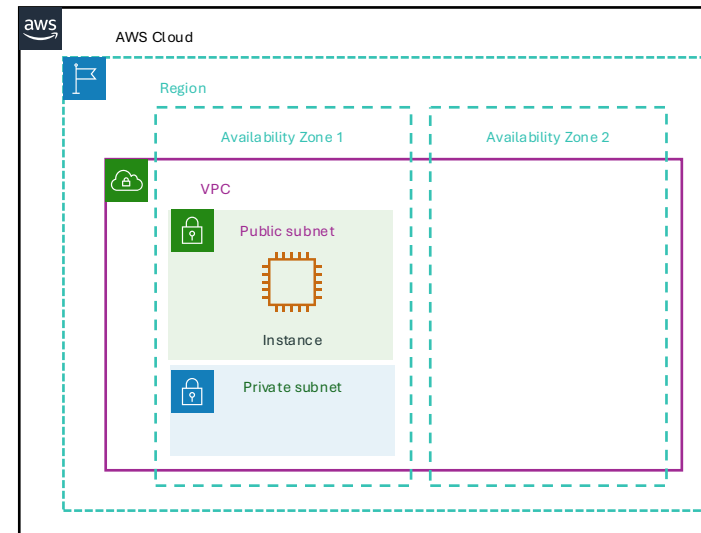
- Estimate the **cost** of a **data pipeline** using the AWS **Pricing Calculator**

✓ Azure Synapse Analytics	①	Tier: Compute Optimized Gen2, De...	Upfront: \$0.00	Monthly: \$6,418.94
✓ Azure Analysis Services	①	Developer (Hours), 1 Instance(s), 72...	Upfront: \$0.00	Monthly: \$95.04
✓ Power BI Embedded	①	1 node(s) x 720 Hours, Node type: ...	Upfront: \$0.00	Monthly: \$1,445.83
✓ Storage Accounts	①	Data Lake Storage Gen2, Standard, ...	Upfront: \$0.00	Monthly: \$15,818...
✓ Azure Cosmos DB	①	Standard provisioned throughput (...)	Upfront: \$0.00	Monthly: \$273.36



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%**

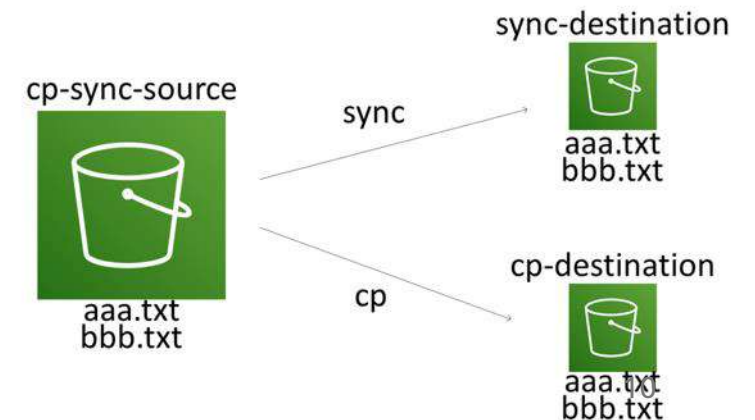
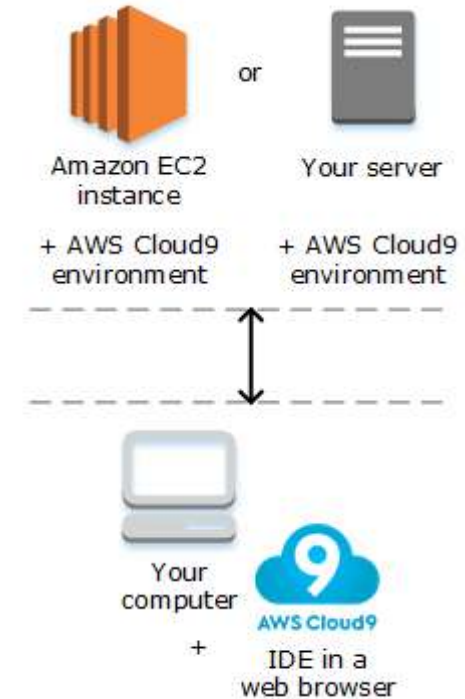
Distributed Storage: Activities

1 HDFS Fault Tolerance

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

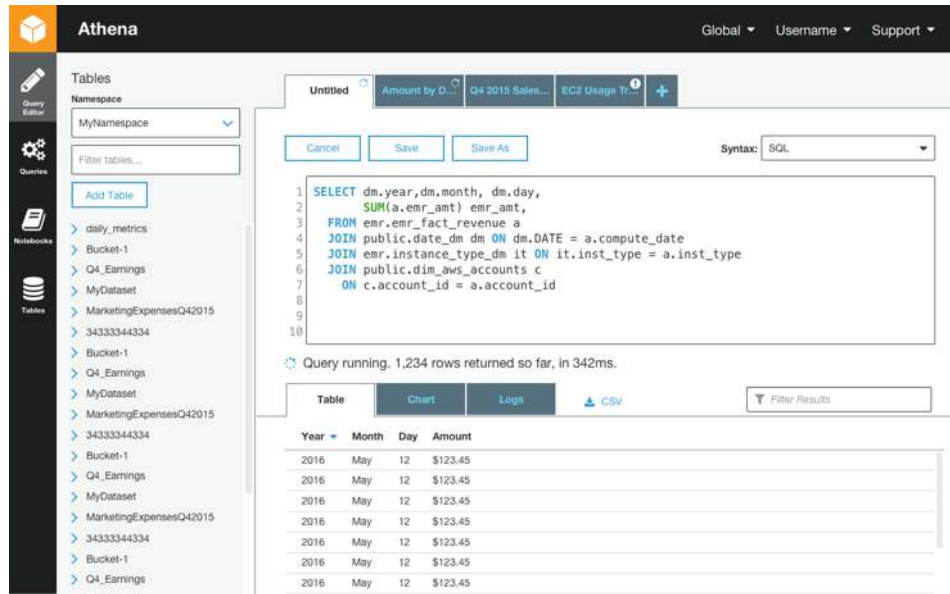
2 Amazon S3 Exploration

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



Distributed Storage: Activities

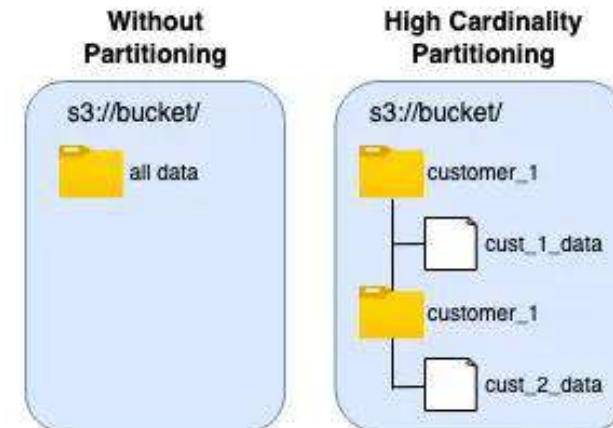
- Demo: Federated queries on Amazon S3 using **Athena**
 - Pricing is based on data scanned (GB scanned per query)



The screenshot shows the Amazon Athena console interface. On the left, there's a sidebar with navigation options like 'Query Editor', 'Tables', and 'Queries'. The main area displays a SQL query in a text editor, which is a SELECT statement joining several tables. Below the query editor, there's a status bar indicating 'Query running. 1,234 rows returned so far, in 342ms.' At the bottom, a table of results is shown with columns 'Year', 'Month', 'Day', and 'Amount'.

```
1 SELECT dm.year, dm.month, dm.day,
2       SUM(a.emr_amt) emr_amt,
3 FROM   emr.emr_fact_revenue a
4 JOIN   public.date_dm dm ON dm.DATE = a.compute_date
5 JOIN   emr.instance_type_dm it ON it.inst_type = a.inst_type
6 JOIN   public.dim_aws_accounts c
7       ON c.account_id = a.account_id
```

Year	Month	Day	Amount
2016	May	12	\$123.45
2016	May	12	\$123.45
2016	May	12	\$123.45
2016	May	12	\$123.45
2016	May	12	\$123.45
2016	May	12	\$123.45
2016	May	12	\$123.45
2016	May	12	\$123.45



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

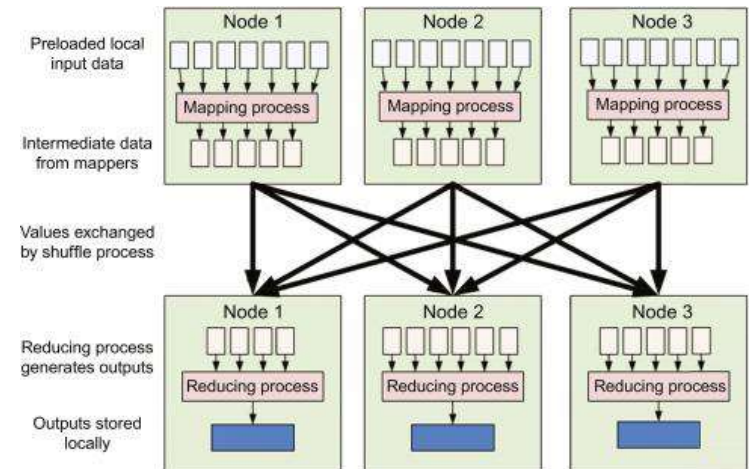
Understanding MapReduce

- 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



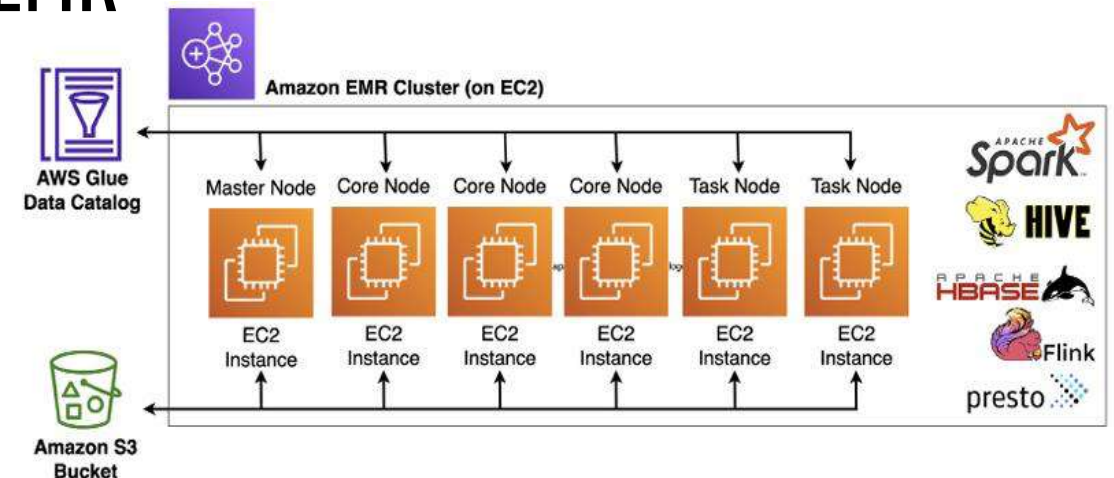
Distributed Processing: Activities

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

Analytics

Processing

Storage

Infrastructure

- 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



Big Data Analytics

About Staff Syllabus Exercises **Datathon** Demofest Resources

1 Analysis of Web Archives from the Digital Humanities

Datathon

- Description
- Instructions
- Evaluation
- Resources
- ★ Queries TODO
- Teams

INTERNET ARCHIVE

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

- 2** Common Crawl maintains a **free, open repository** of web crawl data that can be used by anyone.

Access to the corpus hosted by Amazon is **free**.

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

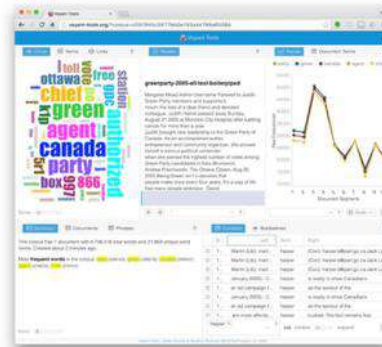
- Extract ALL urls
- Extract top level-domains

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

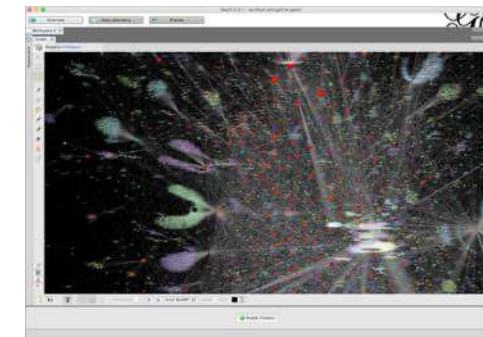
Text Analysis

- Extract text in X language
- Extract text without boilerplate



Link Analysis

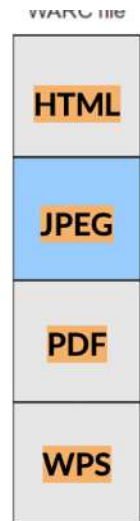
- Extract link structure
- Export to Gephi



AUT tools

1

WARC



WARC record

Text header

Content
block

[image/jpeg binary data]

NB Server



Node

Node

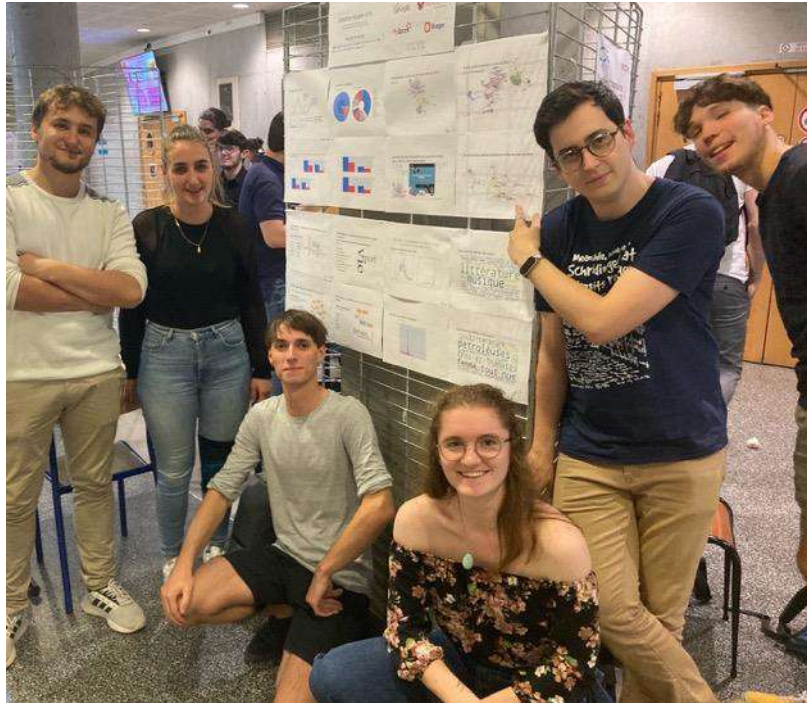
Node

Datalake

WARC



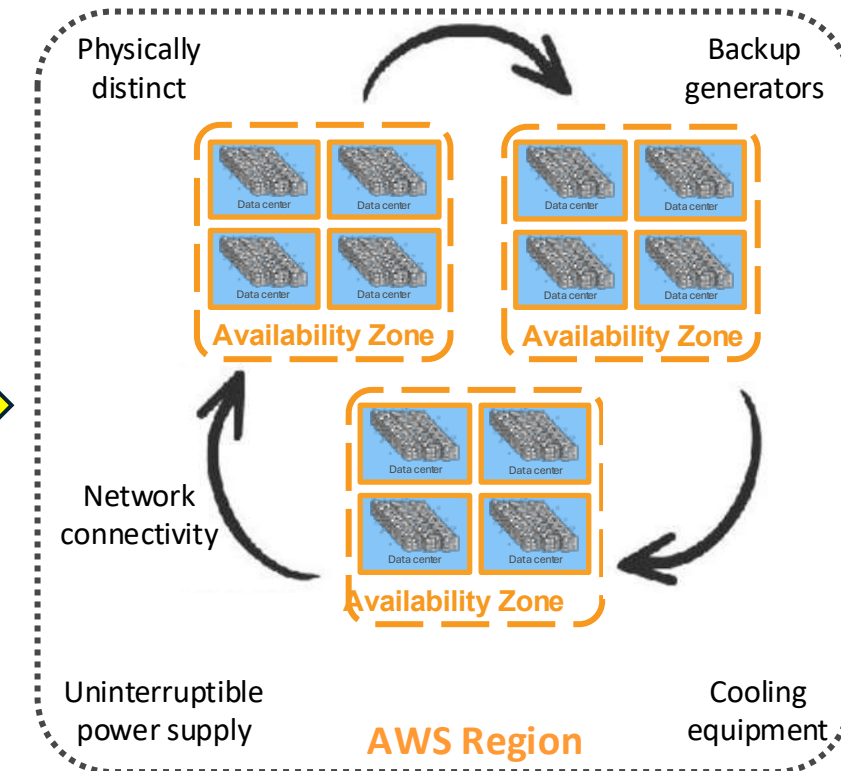
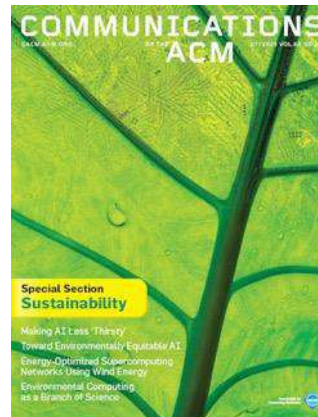
Big Data Analytics: DemoFest



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**



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

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

ABSTRACT

80S GBox is Atanaka's novelized data integration cloud service that can be simple and cost effective to extract, clean, merge, load, and organize data. Originally named GBox 2007, 80S GBox began as an extract transform load (ETL) service designed to reduce data loads and data ingestion of the underutilized heavy IBM mainframe to local databases, data warehouses, and data lakes on Amazon S3. Since then, it has evolved to serve a larger audience including ETL, specialists and data scientists, and include a broader range of data integration capabilities. Tens of thousands of customers use 80S GBox every month.

In this paper, we describe the new costs and challenges cloud customers face in preparing data for their data science projects. We also discuss the challenges we have faced on our journey to deliver data science projects.

[illegible]

ABSTRACT

80S GBox is Atanaka's novelized data integration cloud service that can be simple and cost effective to extract, clean, merge, load, and organize data. Originally named GBox 2007, 80S GBox began as an extract transform load (ETL) service designed to reduce data loads and data ingestion of the underutilized heavy IBM mainframe to local databases, data warehouses, and data lakes on Amazon S3. Since then, it has evolved to serve a larger audience including ETL, specialists and data scientists, and include a broader range of data integration capabilities. Tens of thousands of customers use 80S GBox every month.

In this paper, we describe the new costs and challenges cloud customers face in preparing data for their data science projects. We also discuss the challenges we have faced on our journey to deliver data science projects.

James Bernhoff Amazon Web Services & The University of Toronto at Aurora	Rajeev Joshi Amazon Web Services	Vytautas Antanasukas ETD Zurich
Brendan Cully Amazon Web Services	Bernhard Kragl Amazon Web Services	Seth Muecke Amazon Web Services
Kyle Sauri Amazon Web Services	Drew Schlett Amazon Web Services	Grant Slatten Amazon Web Services
Serdar Tazari	Jacob Van Geffen	Andrew Wargield

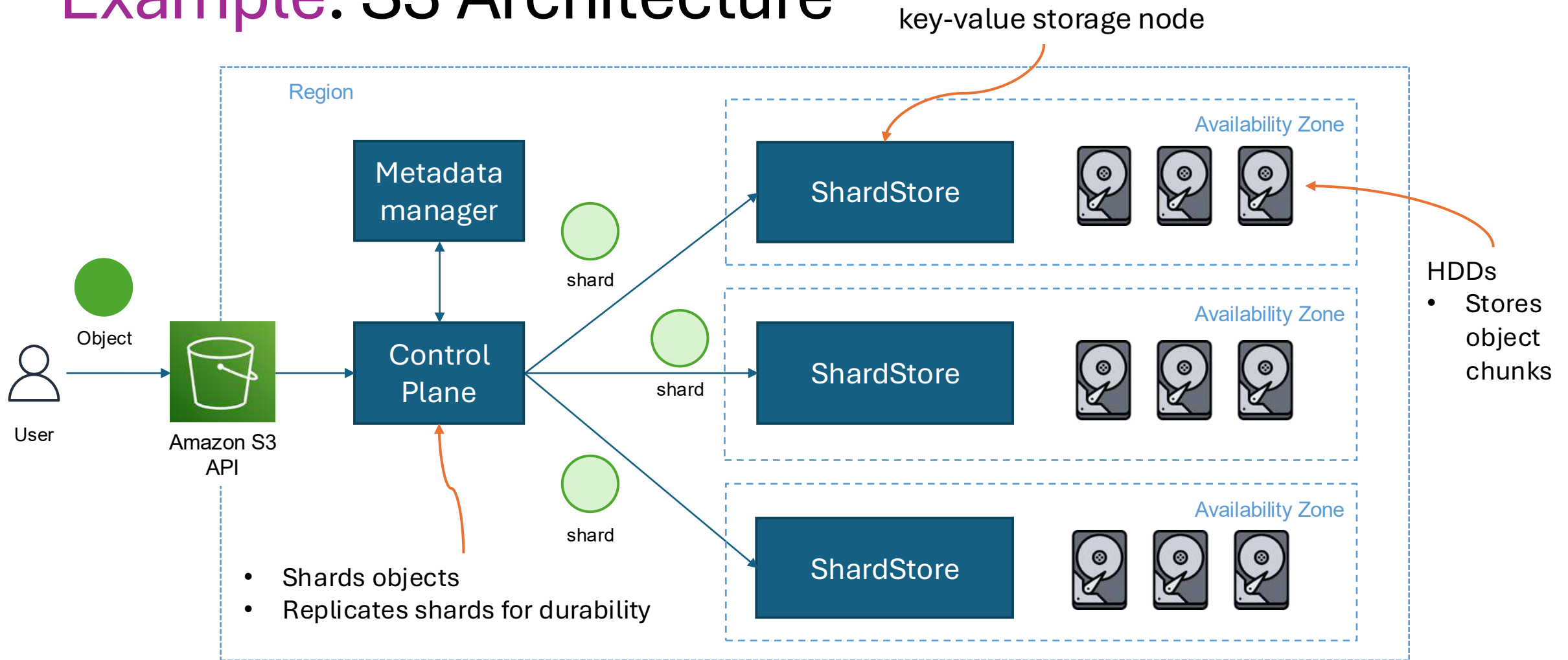
[illegible]

uniformity at both software and its specification evolve over time. Our approach decomposes correctness into independent properties, each checked by the most appropriate tool, and develops verifiable reference models for specifications. In the next section, we apply the design methodology presented in Section 3 to a matching problem, including both the state consistency and concurrency problems, and has been extended by our formal methods experts to check new features and properties in ShellSorter. We conclude.

uniformity at both software and its specification evolve over time. Our approach decomposes correctness into independent properties, each checked by the most appropriate tool, and develops verifiable reference models for specifications. In the next section, we apply the design methodology presented in Section 3 to a matching problem, including both the state consistency and concurrency problems, and has been extended by our formal methods experts to check new features and properties in ShellSorter. We conclude.

[illegible]

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)

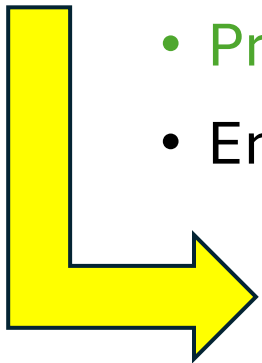
Conclusions

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

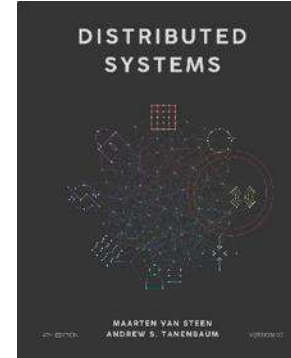
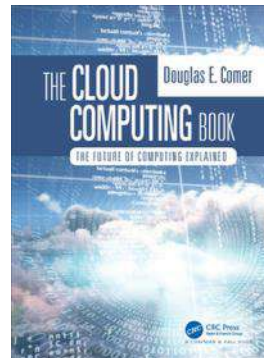
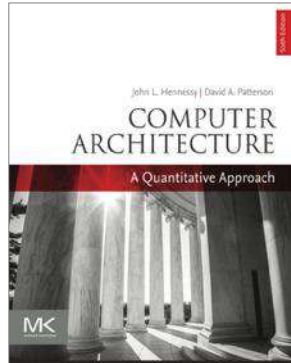
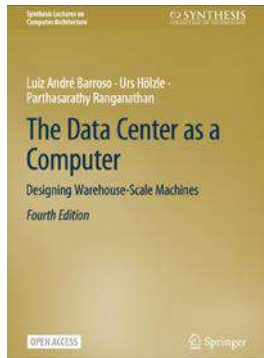
- **Prefer free-tier** or **always-free services**
- Ensure exercises remain viable even if educational programs change



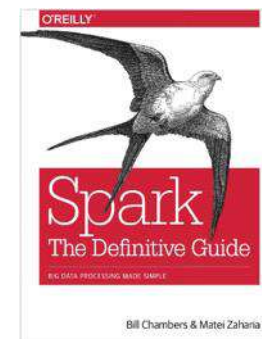
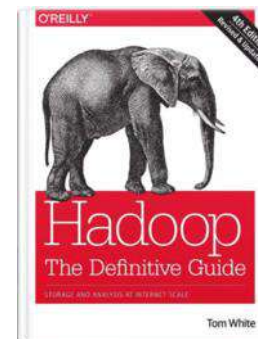
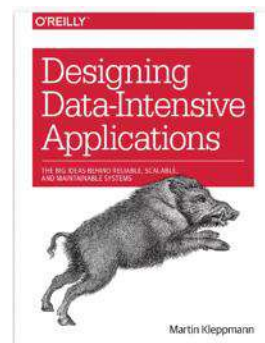
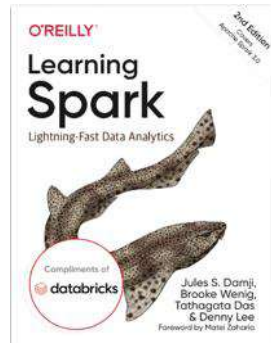
- GCP temporarily suspended educational credits outside the US in 2025
- GitHub Education reduced its Codespaces offering in July 2026

Recommended Resources

Free

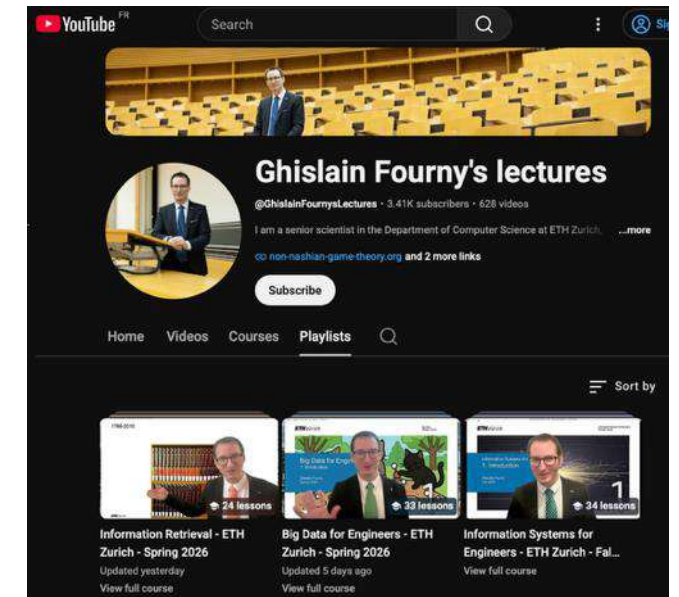


Free



Lectures

- @GhislainFournysLectures



Free

Free (formerly)

Merci
Questions ?