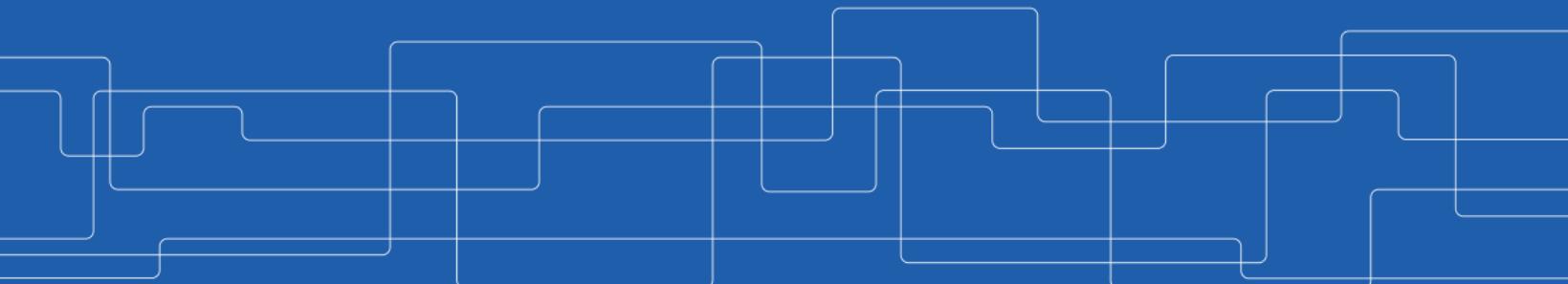




# Foundations of Data-Intensive Computing and Cloud Platforms

Amir H. Payberah  
[payberah@kth.se](mailto:payberah@kth.se)  
2025-08-27





# Course Information



## Course Objective

- ▶ Provide students with a solid foundation for **understanding large scale distributed systems** used for **storing and processing** massive data.
- ▶ Cover a wide variety of advanced topics in **data intensive computing platforms**, i.e., the frameworks to **store and process** big data.



## The Dual Lens of This Course

- ▶ Technical lens
- ▶ Critical lens



# The Dual Lens of This Course

- ▶ Technical lens
  - Understand architectures
  - Learn to design and deploy
- ▶ Critical lens



# The Dual Lens of This Course

## ► Technical lens

- Understand architectures
- Learn to design and deploy

## ► Critical lens

- Who controls and benefits from this infrastructure?
- Reveal hidden labor
- Track environmental costs: energy, water, e-waste



# Why the Critical Lens Matters



Technology is not neutral.

Godfrey Reggio



## Intended Learning Outcomes (ILOs)

- ▶ ILO1: Understand the main concepts of data-intensive computation platforms.



## Intended Learning Outcomes (ILOs)

- ▶ ILO1: Understand the main concepts of data-intensive computation platforms.
- ▶ ILO2: Apply the grabbed knowledge to store and process massive data.



## Intended Learning Outcomes (ILOs)

- ▶ ILO1: Understand the main concepts of data-intensive computation platforms.
- ▶ ILO2: Apply the grabbed knowledge to store and process massive data.
- ▶ ILO3: Analyze the technical merits of data-intensive computation platforms.



## Intended Learning Outcomes (ILOs)

- ▶ ILO1: Understand the main concepts of data-intensive computation platforms.
- ▶ ILO2: Apply the grabbed knowledge to store and process massive data.
- ▶ ILO3: Analyze the technical merits of data-intensive computation platforms.
- ▶ ILO4: Critically assess the social, environmental, and ethical impacts of data-intensive systems.



# The Course Assessment

- ▶ **Task1:** the [review](#) questions.



# The Course Assessment

- ▶ **Task1:** the [review](#) questions.
- ▶ **Task2:** the [project](#) assignment.



# The Course Assessment

- ▶ **Task1:** the [review](#) questions.
- ▶ **Task2:** the [project](#) assignment.
- ▶ **Task3:** the [essay](#) and the [presentation](#).



# The Course Assessment

- ▶ **Task1:** the [review](#) questions.
- ▶ **Task2:** the [project](#) assignment.
- ▶ **Task3:** the [essay](#) and the [presentation](#).
- ▶ **Task4:** the final exam.



# The Course Assessment

- ▶ **Task1:** the **review** questions.
- ▶ **Task2:** the **project** assignment.
- ▶ **Task3:** the **essay** and the **presentation**.
- ▶ **Task4:** the final exam.
- ▶ Task1, Task2, and Task3 should be done in groups of **two/three** students.



## How Each ILO is Assessed?

	Task1	Task2	Task3	Task4
ILO1	x			x
ILO2		x		
ILO3			x	
ILO4	x	x	x	



## Task1: The Review Questions (P/F)

- ▶ Five set of review questions, one set for each week.



## Task2: The Project Assignment (P/F)

- ▶ Proposed by students and confirmed by the teacher.
- ▶ Source code and oral presentation.



## Task3: The Essay and The Presentation (A-F)

- ▶ Selecting **three papers**, writing an **essay** and **presenting** it to their **opponents** (another group).



## Task3: The Essay and The Presentation (A-F)

- ▶ Selecting **three papers**, writing an **essay** and **presenting** it to their **opponents** (another group).
- ▶ Grading of this task has the following parts:



## Task3: The Essay and The Presentation (A-F)

- ▶ Selecting **three papers**, writing an **essay** and **presenting** it to their **opponents** (another group).
- ▶ Grading of this task has the following parts:
  - *E*: **Essay** (5 points)



## Task3: The Essay and The Presentation (A-F)

- ▶ Selecting **three papers**, writing an **essay** and **presenting** it to their **opponents** (another group).
- ▶ Grading of this task has the following parts:
  - *E*: **Essay** (5 points)
  - *P*: **Presentation** (2 points)



## Task3: The Essay and The Presentation (A-F)

- ▶ Selecting **three papers**, writing an **essay** and **presenting** it to their **opponents** (another group).
- ▶ Grading of this task has the following parts:
  - *E*: **Essay** (5 points)
  - *P*: **Presentation** (2 points)
  - *Q*: **Reviewing essay and asking questions** (2 points)



## Task3: The Essay and The Presentation (A-F)

- ▶ Selecting **three papers**, writing an **essay** and **presenting** it to their **opponents** (another group).
- ▶ Grading of this task has the following parts:
  - *E*: **Essay** (5 points)
  - *P*: **Presentation** (2 points)
  - *Q*: **Reviewing essay and asking questions** (2 points)
  - *A*: **Answering questions** (1 point)



## Task3: The Essay and The Presentation (A-F)

- ▶ Selecting **three papers**, writing an **essay** and **presenting** it to their **opponents** (another group).
- ▶ Grading of this task has the following parts:
  - *E*: **Essay** (5 points)
  - *P*: **Presentation** (2 points)
  - *Q*: **Reviewing essay and asking questions** (2 points)
  - *A*: **Answering questions** (1 point)
- ▶ The final grade: A: 10, B: 9, C: 8, D: 7, E: 6, F: <5.



## Task4: The Final Exam (A-F)

- ▶ The final exam covers all the modules presented during the course

# The Final Grade

- ▶ **LAB1 - Programming Assignments**, 3.0 credits (P/F)
  - You must successfully complete both Task1 and Task2.
- ▶ **TEN1 - Examination**, 4.5 credits (A-F)
  - The grade is calculated as  $0.3 \times \text{Task3} + 0.7 \times \text{Task4}$ .
  - You must achieve at least an E in each task to pass.

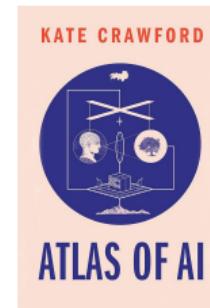
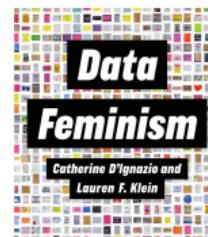
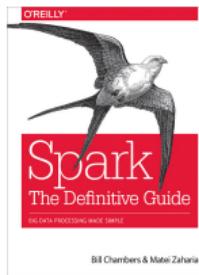
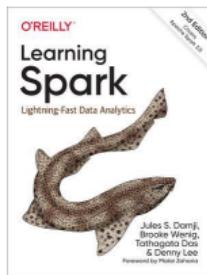
© Randy Glasbergen  
[www.glasbergen.com](http://www.glasbergen.com)



"Why is an A or B better than a C or D?  
Aren't all letters equal in the eyes of God?"

# The Course Material

- ▶ Mainly based on research papers.
- ▶ We also cover the following books.





## The Course Web Page

<https://canvas.kth.se/courses/56249>



# The Course Overview



# Cloud Computing and Big Data

- ▶ The main trends:
  - Computers not getting any faster
  - Internet connections getting faster
  - More people connected to the Internet



# Cloud Computing and Big Data

- ▶ The main trends:
  - Computers not getting any faster
  - Internet connections getting faster
  - More people connected to the Internet
- ▶ Conclusion: move the computation and storage of big data to the cloud!



# Cloud Computing



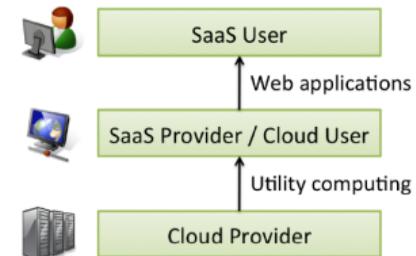
## What is Cloud Computing?

<https://tinyurl.com/5chh94av>



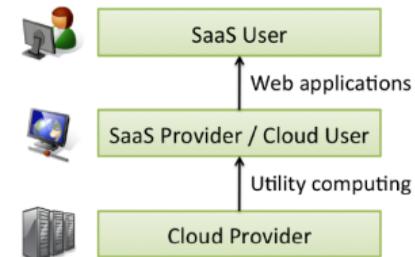
# Cloud Computing Definition

- ▶ Cloud Computing refers to both:
  1. The **applications** delivered as **services** over the Internet
  2. The **hardware and systems software** in the datacenters that provide those **services**



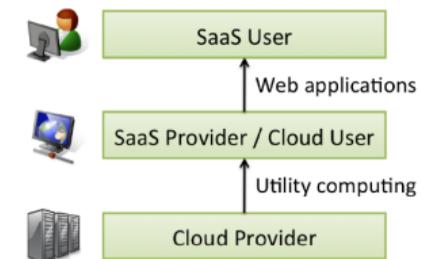
# Cloud Computing Definition

- ▶ Cloud Computing refers to both:
  1. The **applications** delivered as **services** over the Internet
  2. The **hardware and systems software** in the datacenters that provide those **services**
- ▶ The **services**: called **Software as a Service (SaaS)**



# Cloud Computing Definition

- ▶ Cloud Computing refers to both:
  1. The **applications** delivered as **services** over the Internet
  2. The **hardware and systems software** in the datacenters that provide those **services**
- ▶ The **services**: called **Software as a Service (SaaS)**
- ▶ The datacenter **hardware and software** is called **cloud**





- ▶ The NIST definition:
  - Five characteristics
  - Three service models
  - Four deployment models





# Cloud Characteristics

# Cloud Characteristics



On-demand  
self-service



Ubiquitous  
network  
access



Location  
transparent  
resource  
pooling



Rapid  
elasticity



Measured  
service with  
pay per use

[<http://aka.ms/532>]

## Cloud Characteristics - On-demand Self-Service

- ▶ A consumer can **independently** provision **computing capabilities** without **human interaction** with the service provider.



On-demand  
self-service



## Cloud Characteristics - Ubiquitous Network Access

- ▶ Available over the **network**
- ▶ Accessed through mobile phones, laptops, ...



Ubiquitous  
network  
access

## Cloud Characteristics - Resource Pooling

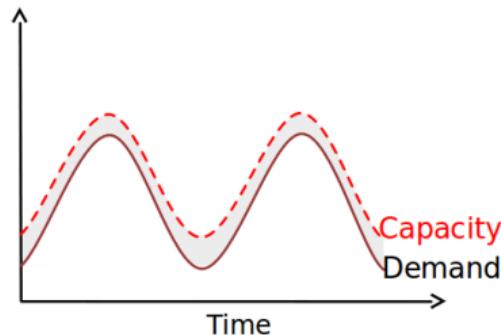
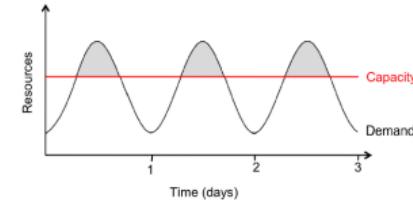
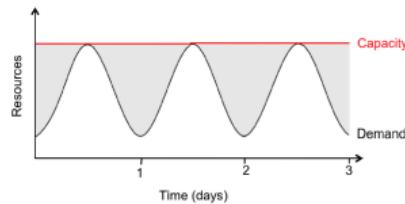
- ▶ Provider's computing resources are **pooled** to serve consumers
- ▶ Location transparent



Location  
transparent  
resource  
pooling

# Cloud Characteristics - Rapid Elasticity

- ▶ Capabilities can be rapidly and **elastically** provisioned, in some cases automatically.



Rapid elasticity

## Cloud Characteristics - Measured Service

- ▶ Resource usage can be monitored, controlled, and reported providing transparency for both the provider and consumer.



Measured  
service with  
pay per use



# Cloud Service Models



# Cloud Service Models



SaaS



PaaS



IaaS

[<http://aka.ms/532>]

- ▶ Assume, you just moved to a city and you are looking for a place to live.



- ▶ What is your choice?



- ▶ What is your choice?
  - Build a [new house](#)?
  - Buy an [empty house](#)?
  - Live in a [hotel](#)?



- ▶ Let's build a **new house!**



- ▶ Let's build a **new house!**
- ▶ You can **fully control** everything you like your new house to have.
- ▶ But that is a **hard work**.



- ▶ What if you buy an [empty house](#)?



- ▶ What if you buy an [empty house](#)?
- ▶ You can [customize](#) some part of your house.
- ▶ But never change the original architecture.



- ▶ How about living in a [hotel](#)?



- ▶ How about living in a **hotel**?
- ▶ Living in a hotel will be a good idea if the only thing you care is about enjoying your life.
- ▶ There is **nothing you can** do with the house except living in it.





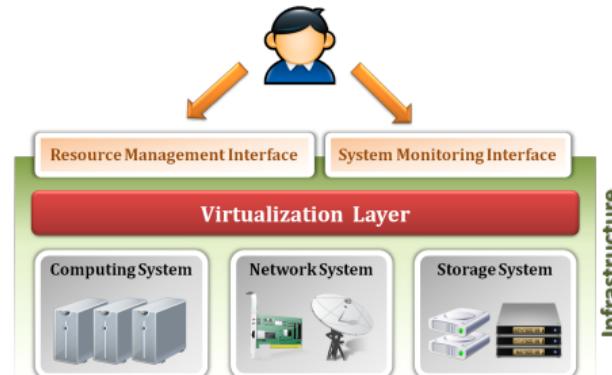
# Let's translate it to Cloud Computing



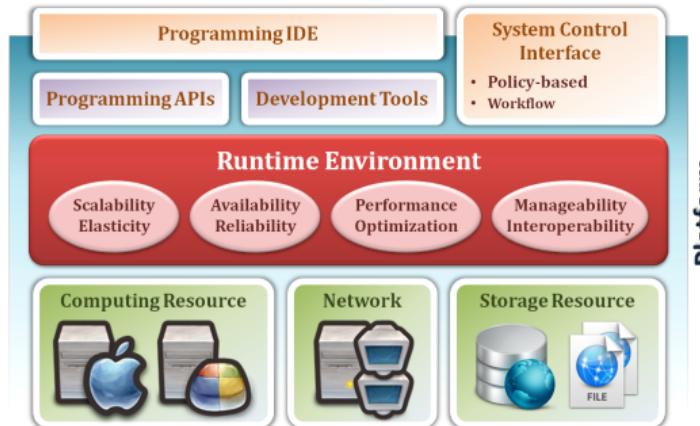
## Service Models

- ▶ Infrastructure as a Service (**IaaS**): similar to **building a new house**.
- ▶ Platform as a Service (**PaaS**): similar to **buying an empty house**.
- ▶ Software as a Service (**SaaS**): similar to **living in a hotel**.

- ▶ Vendor provides **resources**, e.g., processing, storage, network, ...
- ▶ Consumer is provided customized **virtual machines**.
- ▶ Example: Amazon Web Services (EC2 instances and S3 storage)



- ▶ Vendor provides hardware and **development environment**.
- ▶ Example: Google app engine

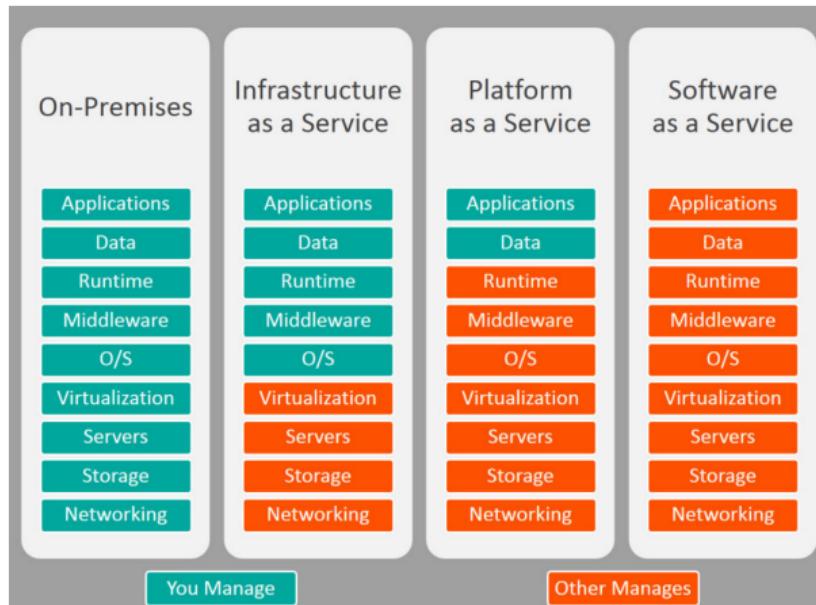


- ▶ Vendor provides **applications** accessed over the network.
- ▶ Example: Gmail, Github





# IaaS - PaaS - SaaS



[<https://goo.gl/xMko1z>]



Go to [www.menti.com](http://www.menti.com), and use the code 4377 6892

► Google App Engine and Netflix?

1. PaaS, IaaS
2. SaaS, PaaS
3. PaaS, SaaS
4. IaaS, PaaS



# Deployment Models

# Deployment Models



**VS**



Publically Shared  
Virtualised Resources

Supports multiple  
customers



Supports connectivity  
over the internet

Suited for less  
confidential information



Privately Shared  
Virtualised Resources

Cluster of dedicated  
customers



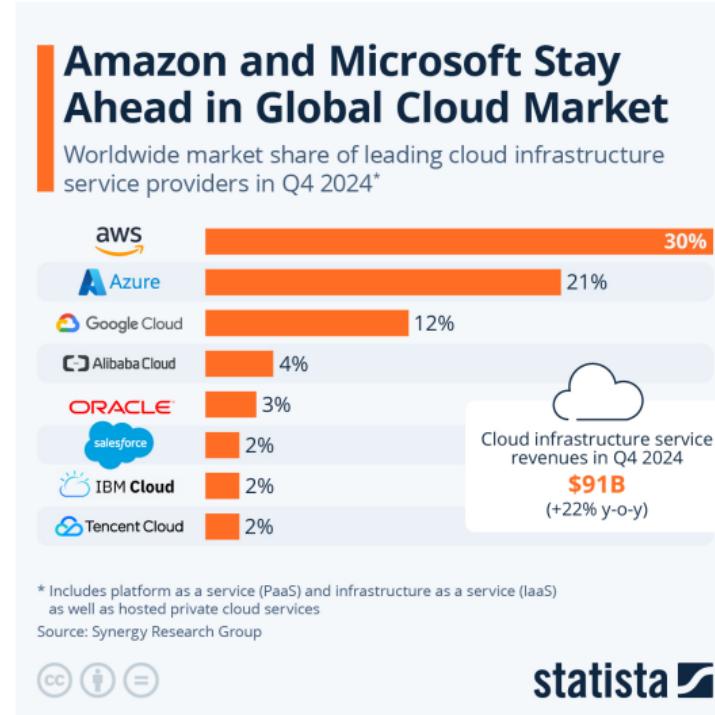
Connectivity over  
internet, fibre and private network



Suited for secured  
confidential information  
& core systems

[<https://goo.gl/fWmcGK>]

# Public Cloud Infrastructure Vendors







## Critical Reflection - Examine Power in the Cloud

- ▶ Who defines the cloud?



## Critical Reflection - Examine Power in the Cloud

- ▶ Who defines the cloud?
- ▶ Which perspectives are missing from these definitions?



## Critical Reflection - Examine Power in the Cloud

- ▶ Who defines the cloud?
- ▶ Which perspectives are missing from these definitions?
- ▶ The definition we used here comes from NIST in the US; how might definitions differ in Europe, Asia, or the Global South?



## Critical Reflection - Examine Power in the Cloud

- ▶ Who defines the cloud?
- ▶ Which perspectives are missing from these definitions?
- ▶ The definition we used here comes from NIST in the US; how might definitions differ in Europe, Asia, or the Global South?
- ▶ Alternative approaches, e.g., GAIA-X, ISO/IEC vs. OpenStack, etc.



# NIST (US) vs. GAIA-X (EU)

NIST (U.S.)	GAIA-X (EU)
<b>Goal:</b> Provide a common language and framework for industry, government, and academia	<b>Goals:</b> Ensure digital sovereignty, build a federated trusted ecosystem, standardize secure cloud operations, stimulate innovation & SME participation
<b>Key Principles:</b> On-demand self-service, broad network access, resource pooling, rapid elasticity, measured service	<b>Key Principles:</b> Transparency & openness, data sovereignty, interoperability & federation, trust & compliance, GDPR alignment, open participation



## Questions

- ▶ Alternative service models?
- ▶ Alternative deployment models?



## Alternative Service Models

- ▶ Federated Data-as-a-Service: community governed data, respect sovereignty.



## Alternative Service Models

- ▶ **Federated Data-as-a-Service:** community governed data, respect sovereignty.
- ▶ **Trust-as-a-Service:** Transparent identity and compliance for fair participation.



## Alternative Service Models

- ▶ **Federated Data-as-a-Service:** community governed data, respect sovereignty.
- ▶ **Trust-as-a-Service:** Transparent identity and compliance for fair participation.
- ▶ **Community-Cloud-as-a-Services:** run by NGOs/universities; access > profit.



## Alternative Service Models

- ▶ **Federated Data-as-a-Service:** community governed data, respect sovereignty.
- ▶ **Trust-as-a-Service:** Transparent identity and compliance for fair participation.
- ▶ **Community-Cloud-as-a-Services:** run by NGOs/universities; access > profit.
- ▶ **Sustainability-as-a-Service:** Report and cap energy, water, carbon use.



## Alternative Service Models

- ▶ **Federated Data-as-a-Service:** community governed data, respect sovereignty.
- ▶ **Trust-as-a-Service:** Transparent identity and compliance for fair participation.
- ▶ **Community-Cloud-as-a-Services:** run by NGOs/universities; access > profit.
- ▶ **Sustainability-as-a-Service:** Report and cap energy, water, carbon use.
- ▶ **Labor-Aware Services:** Make hidden labor visible, ensure fair pay.



# Alternative Deployment Models

- ▶ **Community clouds:** local control, cultural and legal respect.



## Alternative Deployment Models

- ▶ **Community clouds:** local control, cultural and legal respect.
- ▶ **Federated clouds:** shared governance, avoid monopoly power.



# Alternative Deployment Models

- ▶ **Community clouds:** local control, cultural and legal respect.
- ▶ **Federated clouds:** shared governance, avoid monopoly power.
- ▶ **Commons-oriented clouds:** cooperative, nonprofit, public-good infra.



# Alternative Deployment Models

- ▶ **Community clouds:** local control, cultural and legal respect.
- ▶ **Federated clouds:** shared governance, avoid monopoly power.
- ▶ **Commons-oriented clouds:** cooperative, nonprofit, public-good infra.
- ▶ **Green clouds:** renewable-powered, transparent footprint.

# The Environmental Cost of Clouds

- ▶ **Energy consumption:** hyperscale data centers use 200–500 MW (like a mid-size city).



[<https://tinyurl.com/zm9z4y3e>]

# The Environmental Cost of Clouds

- ▶ **Energy consumption:** hyperscale data centers use 200–500 MW (like a mid-size city).
- ▶ **Water for cooling:** millions of liters per day in large facilities.



[<https://tinyurl.com/zm9z4y3e>]

# The Environmental Cost of Clouds

- ▶ **Energy consumption:** hyperscale data centers use 200–500 MW (like a mid-size city).
- ▶ **Water for cooling:** millions of liters per day in large facilities.
- ▶ **Carbon emissions:** 2–3% of global CO<sub>2</sub> emissions from cloud infrastructure.



[<https://tinyurl.com/zm9z4y3e>]

# The Environmental Cost of Clouds

- ▶ **Energy consumption:** hyperscale data centers use 200–500 MW (like a mid-size city).
- ▶ **Water for cooling:** millions of liters per day in large facilities.
- ▶ **Carbon emissions:** 2–3% of global CO<sub>2</sub> emissions from cloud infrastructure.
- ▶ **Mineral extraction:** rare earths, cobalt, and other minerals mined for servers and batteries.



[<https://tinyurl.com/zm9z4y3e>]

# The Invisible Labor Behind the Cloud

- ▶ **Hardware manufacturing:** mining mineral resources, often under dangerous and exploitative conditions.



[<https://tinyurl.com/yt2amtjj>]

# The Invisible Labor Behind the Cloud

- ▶ **Hardware manufacturing:** mining mineral resources, often under dangerous and exploitative conditions.
- ▶ **Data labeling and moderation:** low-paid labeling and content moderation.



[<https://tinyurl.com/yt2amtjj>]

# The Invisible Labor Behind the Cloud

- ▶ **Hardware manufacturing:** mining mineral resources, often under dangerous and exploitative conditions.
- ▶ **Data labeling and moderation:** low-paid labeling and content moderation.
- ▶ **Maintenance and support:** outsourced server upkeep with insecure jobs.



[<https://tinyurl.com/yt2amtjj>]

# The Invisible Labor Behind the Cloud

- ▶ **Hardware manufacturing:** mining mineral resources, often under dangerous and exploitative conditions.
- ▶ **Data labeling and moderation:** low-paid labeling and content moderation.
- ▶ **Maintenance and support:** outsourced server upkeep with insecure jobs.
- ▶ **E-waste processing:** unsafe dismantling of discarded electronics.



[<https://tinyurl.com/yt2amtjj>]



## Questions

- ▶ How can we reduce the **environmental cost** of cloud computing?
- ▶ How can we improve **labor conditions** across the cloud's supply chain?



# Reducing the Environmental Cost of Cloud

- ▶ Use **renewable energy** to power data centers.



# Reducing the Environmental Cost of Cloud

- ▶ Use **renewable energy** to power data centers.
- ▶ Report **energy, water, and carbon** use openly so users see the real cost.



## Reducing the Environmental Cost of Cloud

- ▶ Use **renewable energy** to power data centers.
- ▶ Report **energy, water, and carbon** use openly so users see the real cost.
- ▶ Store **fewer unnecessary** copies of data to save power.



# Reducing the Environmental Cost of Cloud

- ▶ Use **renewable energy** to power data centers.
- ▶ Report **energy, water, and carbon** use openly so users see the real cost.
- ▶ Store **fewer unnecessary** copies of data to save power.
- ▶ Place data centers where **clean energy is available**, not just where it's cheap.



# Reducing the Environmental Cost of Cloud

- ▶ Use **renewable energy** to power data centers.
- ▶ Report **energy, water, and carbon** use openly so users see the real cost.
- ▶ Store **fewer unnecessary** copies of data to save power.
- ▶ Place data centers where **clean energy is available**, not just where it's cheap.
- ▶ Make companies **responsible for pollution** (e.g., carbon taxes).



# Improving Labor Conditions in the Cloud Supply Chain

- ▶ Guarantee living wages for miners, factory workers, and data labelers, not minimum survival pay.



# Improving Labor Conditions in the Cloud Supply Chain

- ▶ Guarantee living wages for miners, factory workers, and data labelers, not minimum survival pay.
- ▶ Ban unsafe practices in mines, factories, and recycling sites; enforce strong global safety standards.



# Improving Labor Conditions in the Cloud Supply Chain

- ▶ Guarantee living wages for miners, factory workers, and data labelers, not minimum survival pay.
- ▶ Ban unsafe practices in mines, factories, and recycling sites; enforce strong global safety standards.
- ▶ Name and credit workers across the supply chain, making their contributions visible instead of hidden.



# Improving Labor Conditions in the Cloud Supply Chain

- ▶ Guarantee living wages for miners, factory workers, and data labelers, not minimum survival pay.
- ▶ Ban unsafe practices in mines, factories, and recycling sites; enforce strong global safety standards.
- ▶ Name and credit workers across the supply chain, making their contributions visible instead of hidden.
- ▶ Give workers a voice in decisions about contracts, sourcing, and fair treatment.



# Improving Labor Conditions in the Cloud Supply Chain

- ▶ Guarantee living wages for miners, factory workers, and data labelers, not minimum survival pay.
- ▶ Ban unsafe practices in mines, factories, and recycling sites; enforce strong global safety standards.
- ▶ Name and credit workers across the supply chain, making their contributions visible instead of hidden.
- ▶ Give workers a voice in decisions about contracts, sourcing, and fair treatment.
- ▶ Cut exploitation at the root by reusing and recycling hardware to reduce demand for new minerals.



# Big Data



# What is Big Data?

<https://tinyurl.com/44b22wzw>



# What is Big Data?



[<https://www.sue-anderson.com.au/index.php/2017/08/18/cursing-curious-work>]



# Big Data

Big data is the data characterized by 4 key attributes: volume, variety, velocity and value.

**ORACLE®**



# Big Data

Big data is the data characterized by 4 key attributes: volume, variety, velocity and value.

Buzzwords

ORACLE®



# Big Data in Simple Words



# Big Data



**DevOps Borat**  
@DEVOPS\_BORAT

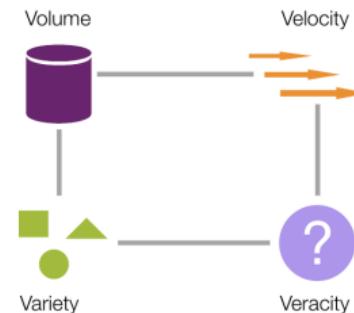
Small Data is when is fit in RAM.  
Big Data is when is crash because  
is not fit in RAM.

2/6/13, 8:22 AM



# The Four Dimensions of Big Data

- ▶ **Volume:** data size
- ▶ **Velocity:** data generation rate
- ▶ **Variety:** data heterogeneity
- ▶ This 4th **V** is for **Vacillation:**  
Veracity/Variability/Value



# One More V

- **Values**, e.g., diversity, equity, inclusion, justice





# How To Store and Process Big Data?



# Problem

- ▶ Traditional platforms **fail** to show the expected performance.
- ▶ Need **new systems** to **store and process** large-scale data

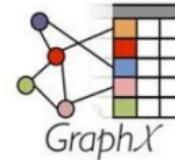
# Scale Up vs. Scale Out

- ▶ Scale **up** or scale **vertically**: adding **resources** to a **single node** in a system.
- ▶ Scale **out** or scale **horizontally**: adding **more nodes** to a system.





APACHE  
**HBASE**



Storm



Spark



S4 distributed stream computing platform



cassandra

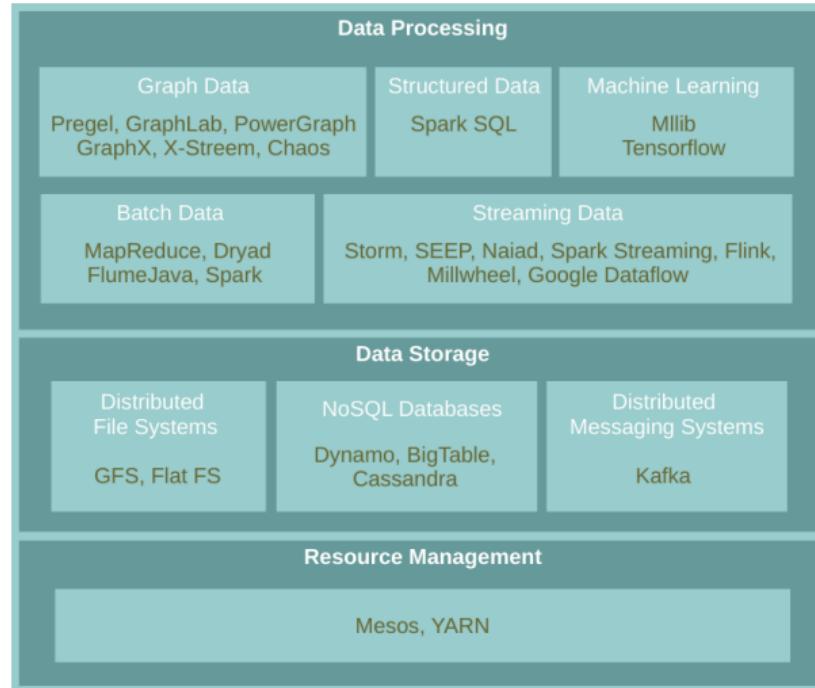


Google Cloud Platform





# Big Data Stack





# Summary



## Summary

- ▶ Cloud computing: characteristics, services, deployments
- ▶ Big data: scale out platforms to store and process
- ▶ Tech is not neutral; it carries the values of their creators
- ▶ Critical consciousness



## References

- ▶ D. Sikeridis et al., A Comparative Taxonomy and Survey of Public Cloud Infrastructure Vendors, arXiv preprint arXiv:1710.01476, 2017.
- ▶ A. Fox et al., Above the clouds: A berkeley view of cloud computing, UCB/EECS 28.13 (2009): 2009.
- ▶ P. Mell et al., The NIST definition of cloud computing, 2011.
- ▶ C. D'ignazio and L. Klein. Data feminism. MIT press, 2023 (Ch1).
- ▶ K. Crawford, The atlas of AI. Yale University Press, 2021 (Ch1-2).



# Questions?