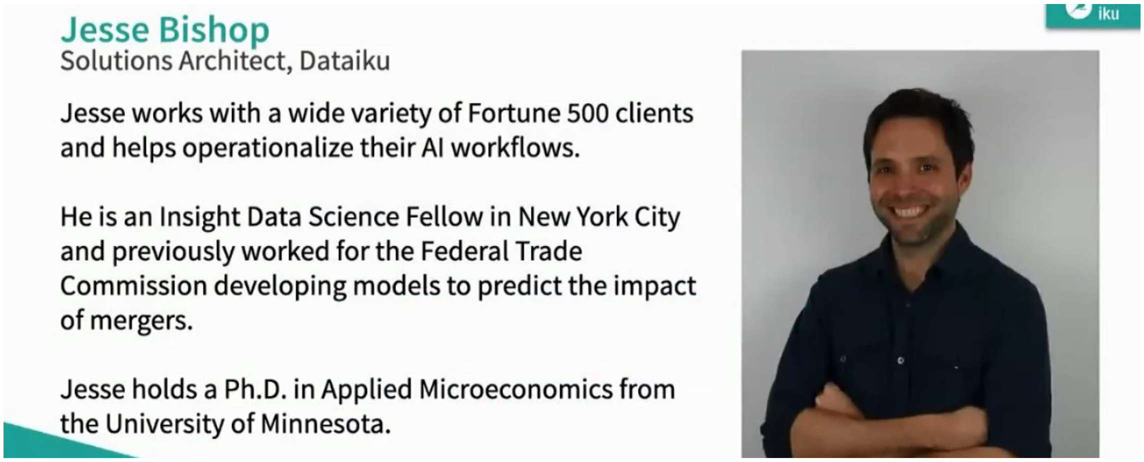
# Crash Course in Data Architecture

Jesse Bishop & Christina Hsiao

Data architecture is the foundation of every organization’s data strategy, but it's not just something for CIOs and data architects - everyone can benefit from understanding the ways data moves between teams and flows into data projects to yield insights. Learn about the key architecture terms and different priorities regarding security and scalability in this crash course.



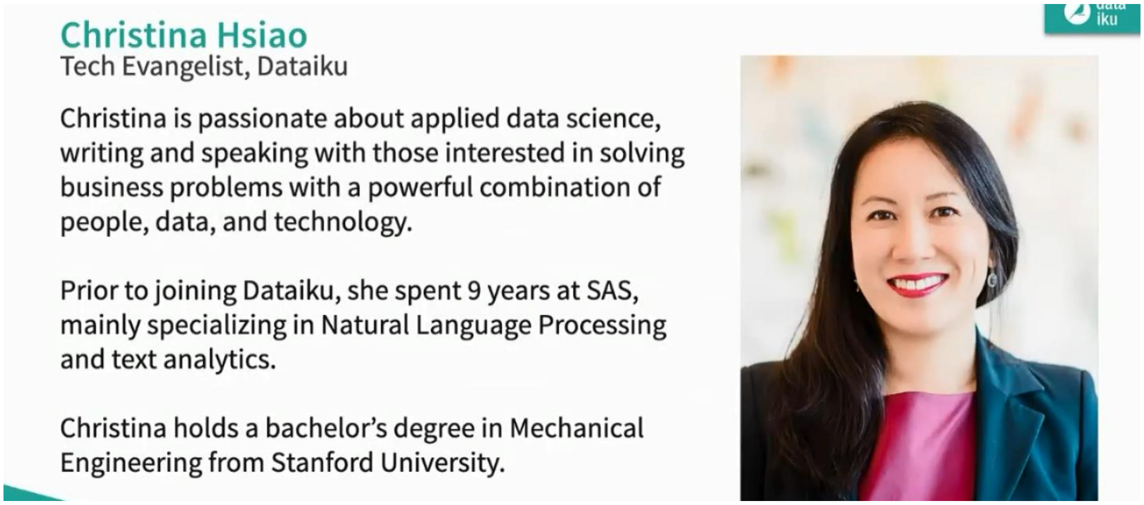
## Jesse Bishop

Solutions Architect, Dataiku

Jesse works with a wide variety of Fortune 500 clients and helps operationalize their AI workflows.

He is an Insight Data Science Fellow in New York City and previously worked for the Federal Trade Commission developing models to predict the impact of mergers.

Jesse holds a Ph.D. in Applied Microeconomics from the University of Minnesota.



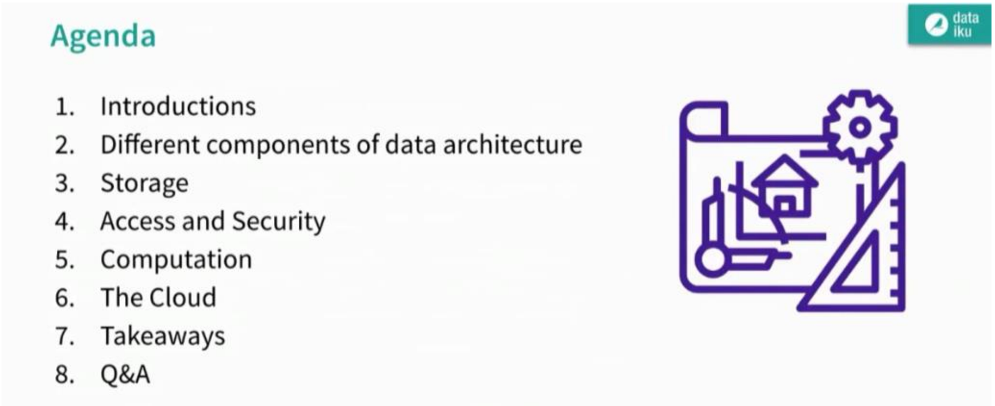
## Christina Hsiao

Tech Evangelist, Dataiku

Christina is passionate about applied data science, writing and speaking with those interested in solving business problems with a powerful combination of people, data, and technology.

Prior to joining Dataiku, she spent 9 years at SAS, mainly specializing in Natural Language Processing and text analytics.

Christina holds a bachelor’s degree in Mechanical Engineering from Stanford University.



## Agenda

1. Introductions
2. Different components of data architecture
3. Storage
4. Access and Security
5. Computation
6. The Cloud
7. Takeaways
8. Q&A

## Data architecture must reflect organizational needs

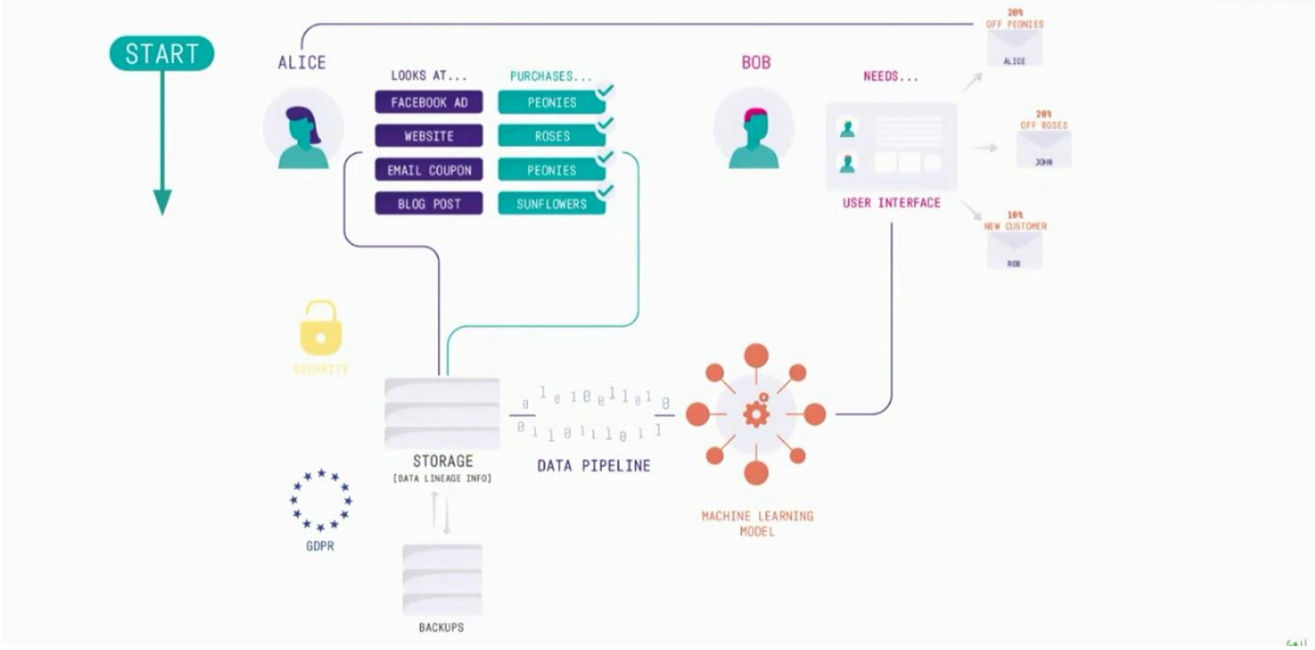
***Challenge***

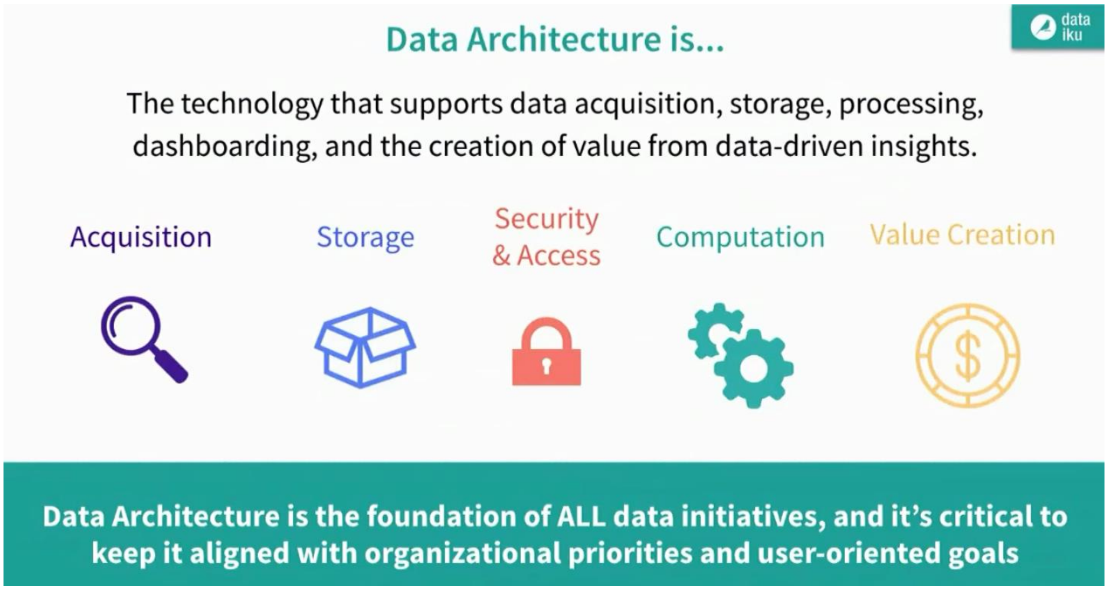
*As an organization scales, can it maintain agility and handle both increased data volumes and data computation demands?*

***Implications***

- *Data is only meaningful & useful when it's being leveraged. Don't store data only for safekeeping; use it to drive decision making*
- *Without a well-planned data architecture that can evolve over time, organizations run the risks of lost or inconsistent data, privacy breaches, and an inability to take advantage of modern data science methods that will keep them relevant and competitive*

## What might Bob's business want to do with data?





## Data Architecture is...

The technology that supports data acquisition, storage, processing, dashboarding, and the creation of value from data-driven insights.

Acquisition

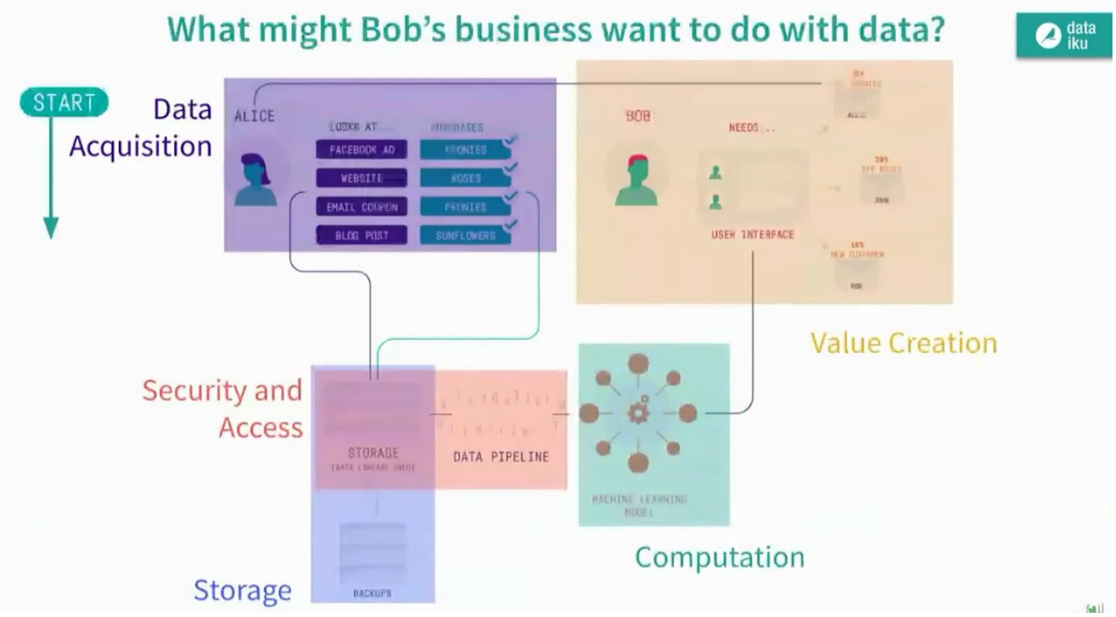
Storage

Security & Access

Computation

Value Creation

**Data Architecture is the foundation of ALL data initiatives, and it's critical to keep it aligned with organizational priorities and user-oriented goals**



## What might Bob's business want to do with data?

START

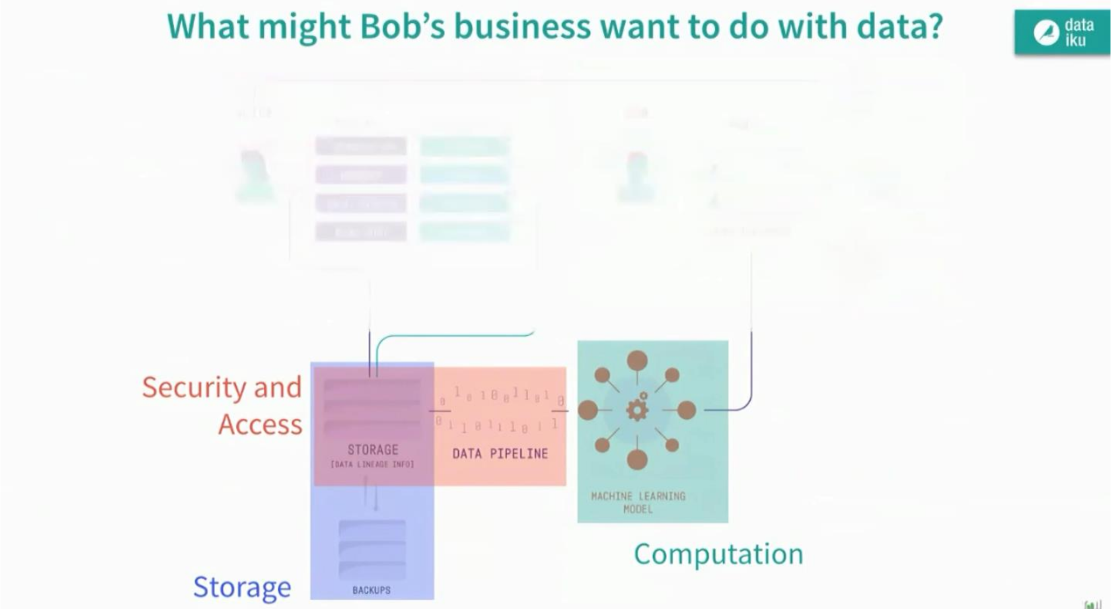
Data Acquisition

Value Creation

Security and Access

Computation

Storage



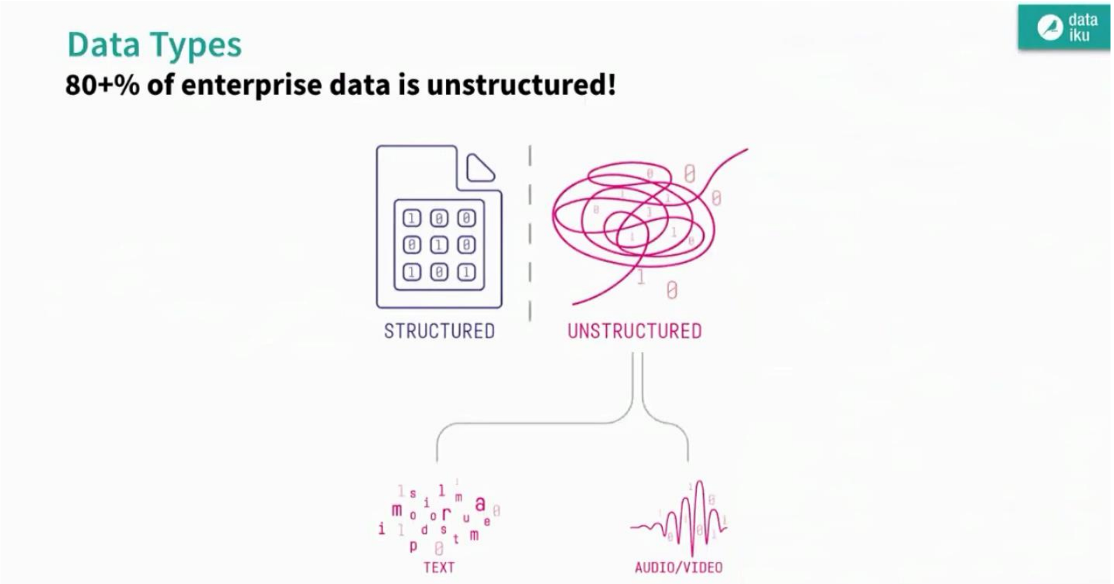
## What might Bob's business want to do with data?

Security and Access

Computation

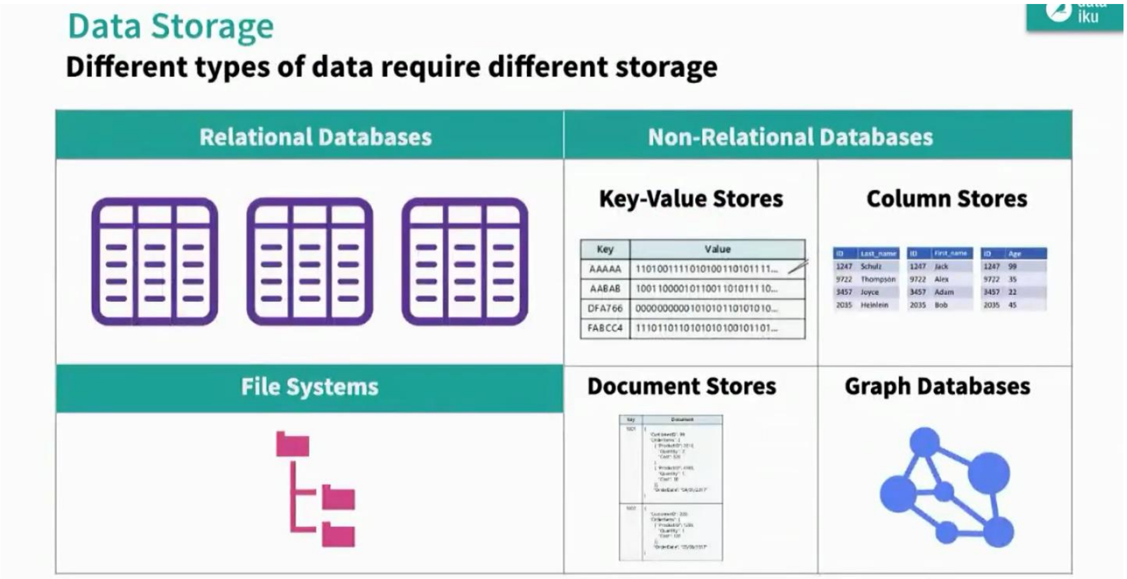
Storage

# Storage



## Data Types

**80+% of enterprise data is unstructured!**



## Data Storage

**Different types of data require different storage**

**Relational Databases**

**Non-Relational Databases**

**Key-Value Stores**

| Key | Value |
|---|---|
| AAAAA | 11010011110101001101011 11... |
| AABAB | 10011000010110011010111 10... |
| DFA766 | 00000000001010101101010 10... |
| FABCC4 | 11101101101010101001011 01... |

**Column Stores**

| ID | Last_name |
|---|---|
| 1247 | Schulz |
| 9722 | Thompson |
| 3457 | Joyce |
| 2035 | Heinlein |

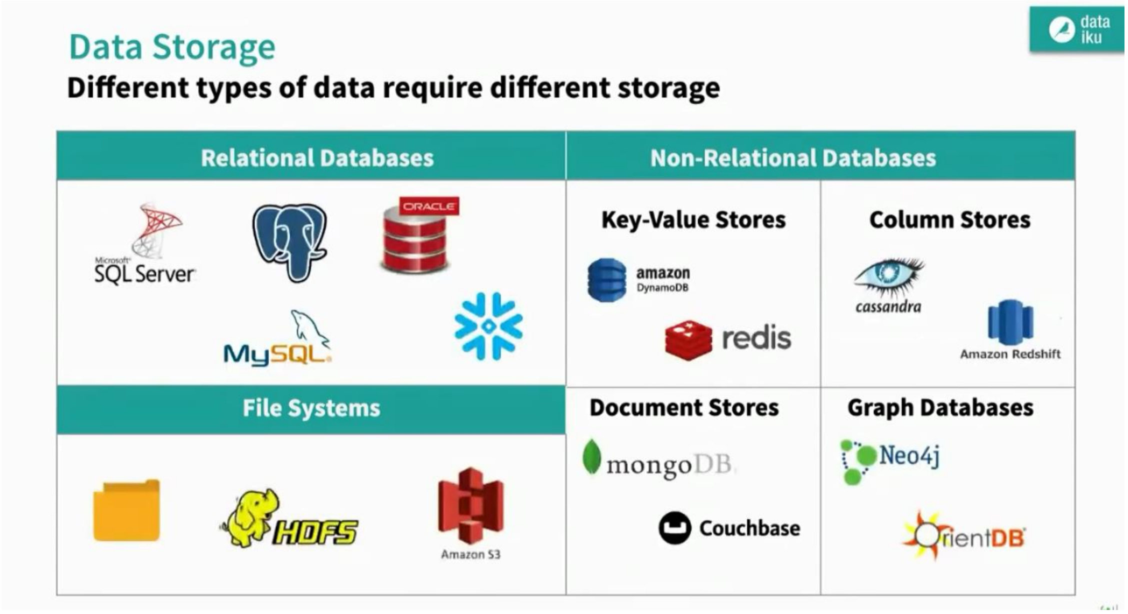
| ID | First_name |
|---|---|
| 1247 | Jack |
| 9722 | Alex |
| 3457 | Adam |
| 2035 | Bob |

| ID | Age |
|---|---|
| 1247 | 99 |
| 9722 | 35 |
| 3457 | 22 |
| 2035 | 45 |

**File Systems**

**Document Stores**

**Graph Databases**

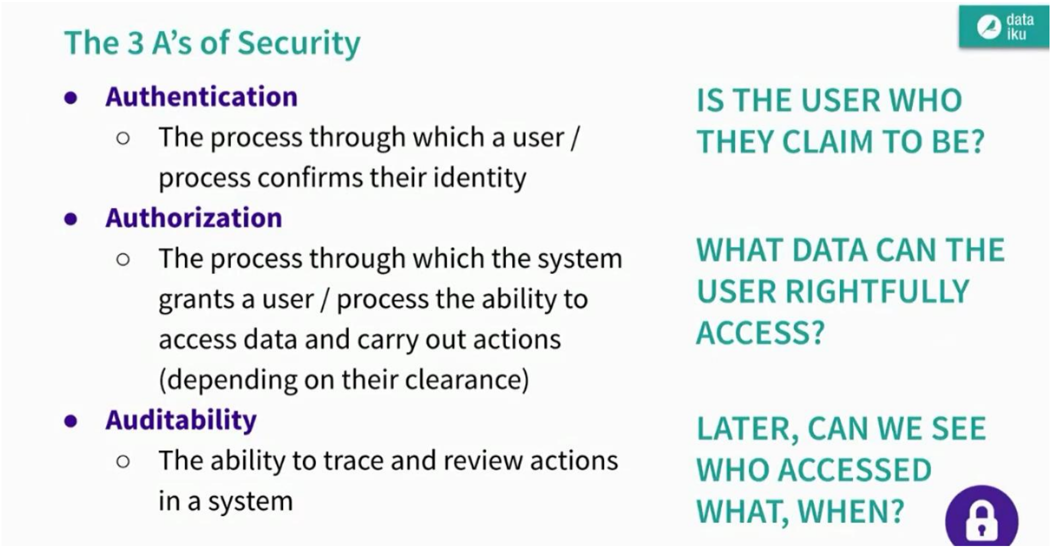


## Data Storage

**Different types of data require different storage**

| Relational Databases | Non-Relational Databases | |
|---|---|---|
| | **Key-Value Stores** | **Column Stores** |
| **File Systems** | **Document Stores** | **Graph Databases** |

# Access & Security



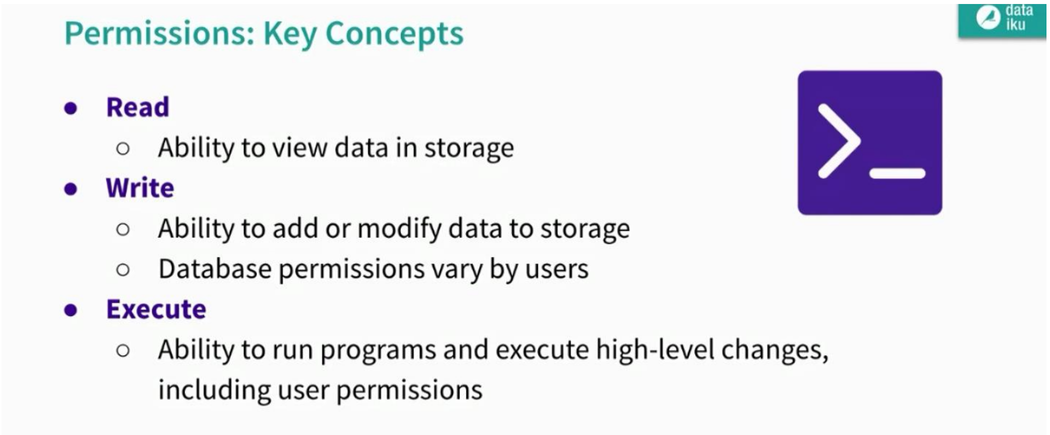
## The 3 A's of Security

- **Authentication**
  - The process through which a user / process confirms their identity
- **Authorization**
  - The process through which the system grants a user / process the ability to access data and carry out actions (depending on their clearance)
- **Auditability**
  - The ability to trace and review actions in a system

IS THE USER WHO THEY CLAIM TO BE?

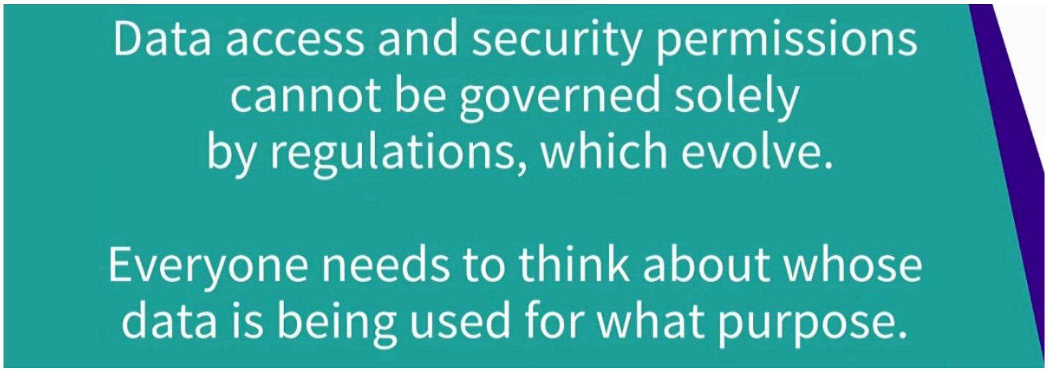
WHAT DATA CAN THE USER RIGHTFULLY ACCESS?

LATER, CAN WE SEE WHO ACCESSED WHAT, WHEN?



## Permissions: Key Concepts

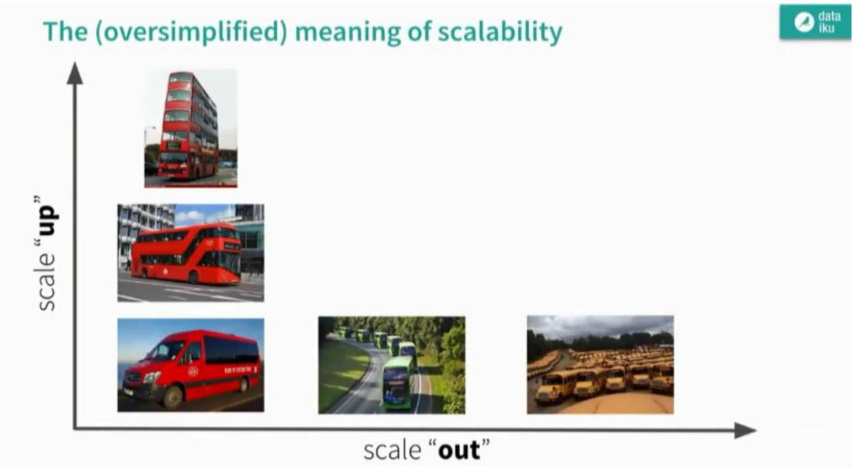
- **Read**
  - Ability to view data in storage
- **Write**
  - Ability to add or modify data to storage
  - Database permissions vary by users
- **Execute**
  - Ability to run programs and execute high-level changes, including user permissions



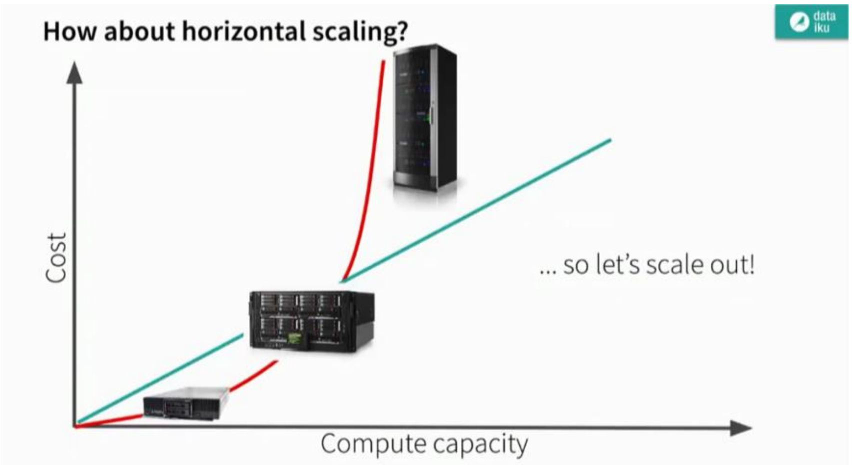
Data access and security permissions cannot be governed solely by regulations, which evolve.

Everyone needs to think about whose data is being used for what purpose.

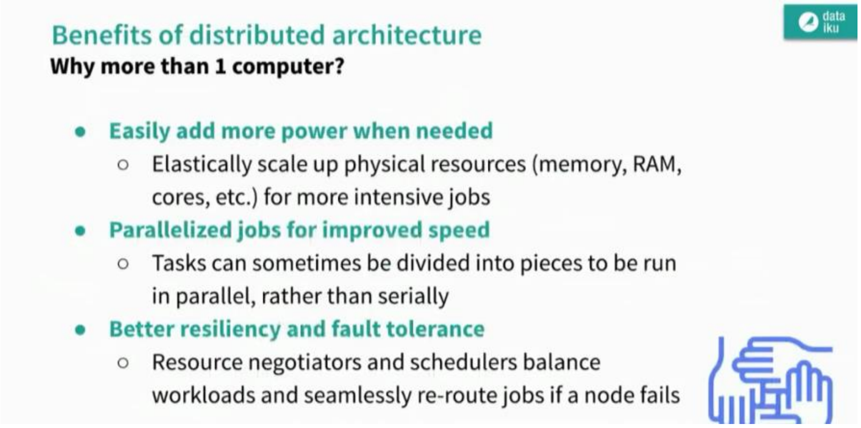
## Computation



### The (oversimplified) meaning of scalability



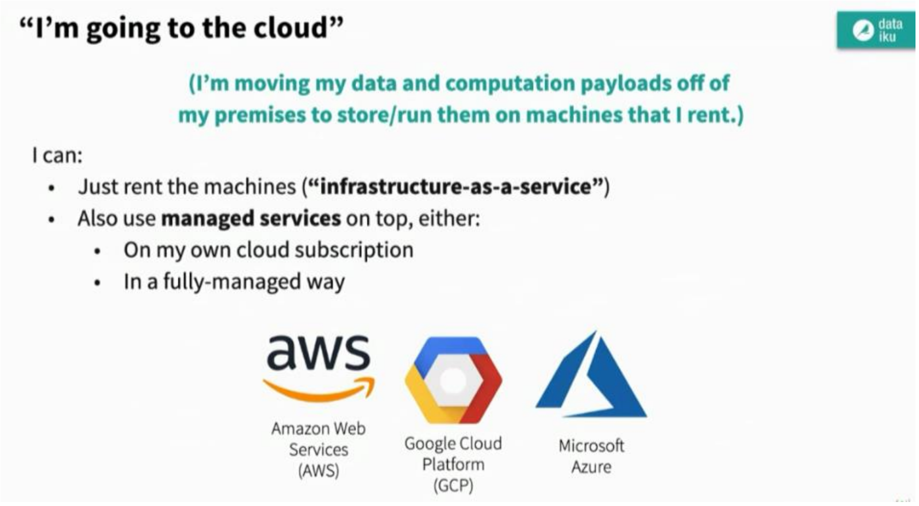
### How about horizontal scaling?



### Benefits of distributed architecture

**Why more than 1 computer?**

- **Easily add more power when needed**
  - Elastically scale up physical resources (memory, RAM, cores, etc.) for more intensive jobs
- **Parallelized jobs for improved speed**
  - Tasks can sometimes be divided into pieces to be run in parallel, rather than serially
- **Better resiliency and fault tolerance**
  - Resource negotiators and schedulers balance workloads and seamlessly re-route jobs if a node fails



## "I'm going to the cloud"

(I'm moving my data and computation payloads off of my premises to store/run them on machines that I rent.)

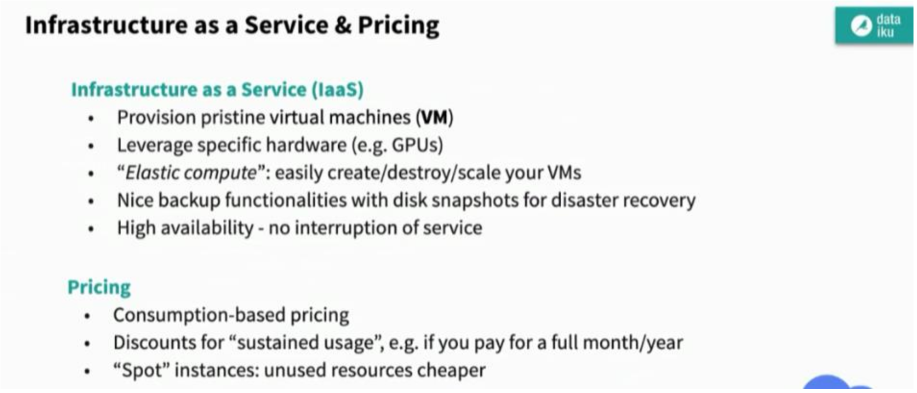
I can:

- Just rent the machines (**"infrastructure-as-a-service"**)
- Also use **managed services** on top, either:
  - On my own cloud subscription
  - In a fully-managed way

Amazon Web Services (AWS)

Google Cloud Platform (GCP)

Microsoft Azure



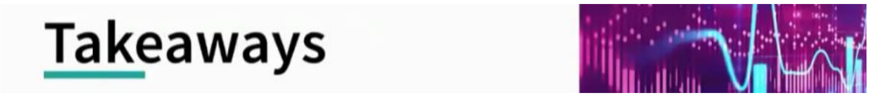
## Infrastructure as a Service & Pricing

### Infrastructure as a Service (IaaS)

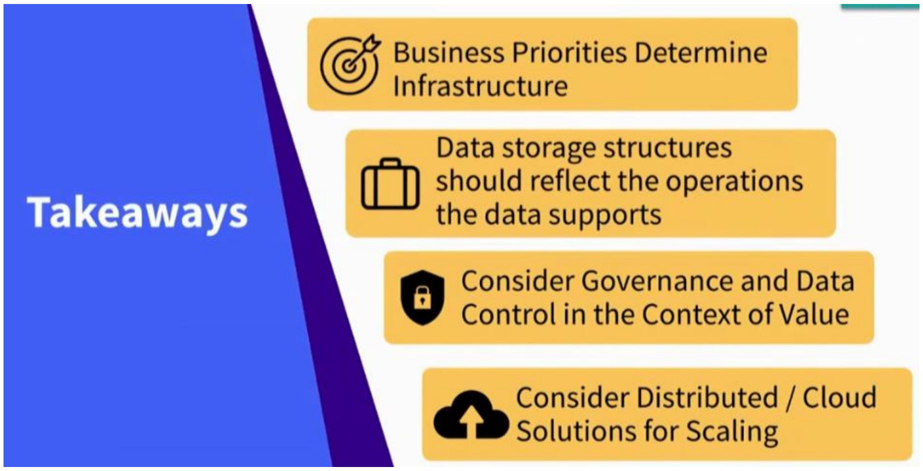
- Provision pristine virtual machines (**VM**)
- Leverage specific hardware (e.g. GPUs)
- "*Elastic compute*": easily create/destroy/scale your VMs
- Nice backup functionalities with disk snapshots for disaster recovery
- High availability - no interruption of service

### Pricing

- Consumption-based pricing
- Discounts for "sustained usage", e.g. if you pay for a full month/year
- "Spot" instances: unused resources cheaper



# Takeaways



## Takeaways

- Business Priorities Determine Infrastructure
- Data storage structures should reflect the operations the data supports
- Consider Governance and Data Control in the Context of Value
- Consider Distributed / Cloud Solutions for Scaling