PLATFORMS TO SUPPORT EMERGING DATA

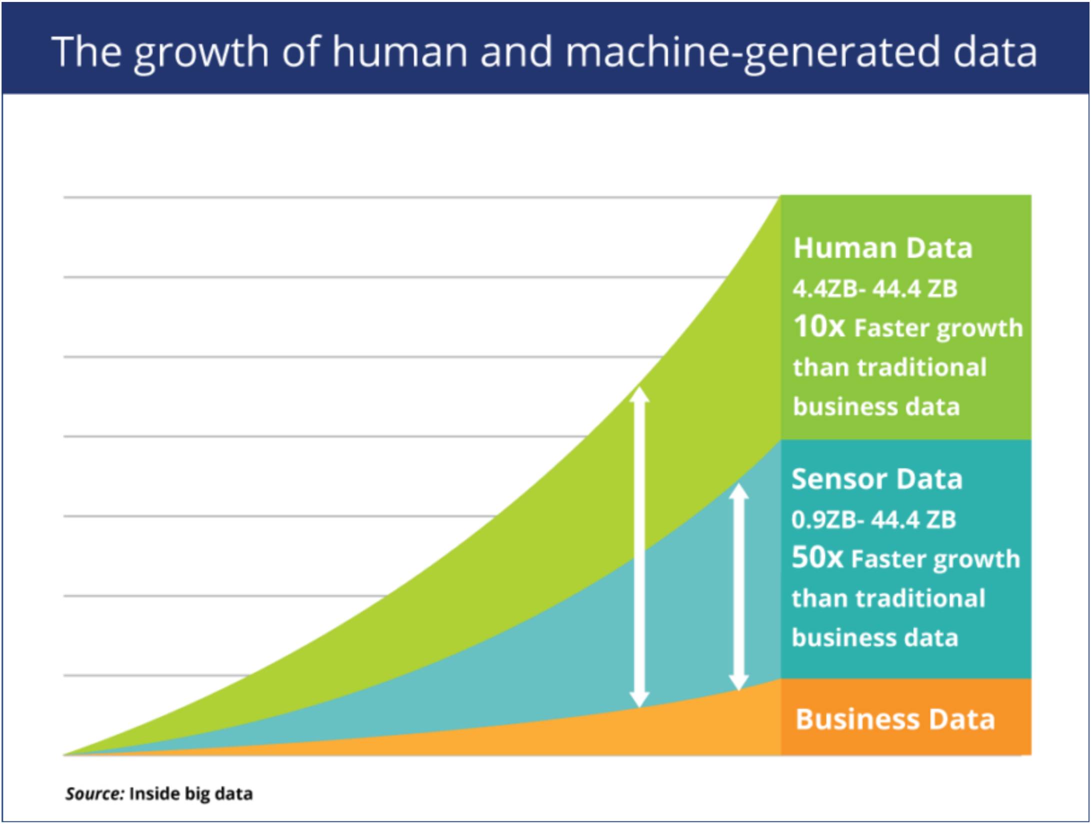
Benjamin Pecheux

AEM Corporation / FHWA EDC-6

Crowdsourcing Contract Support Team

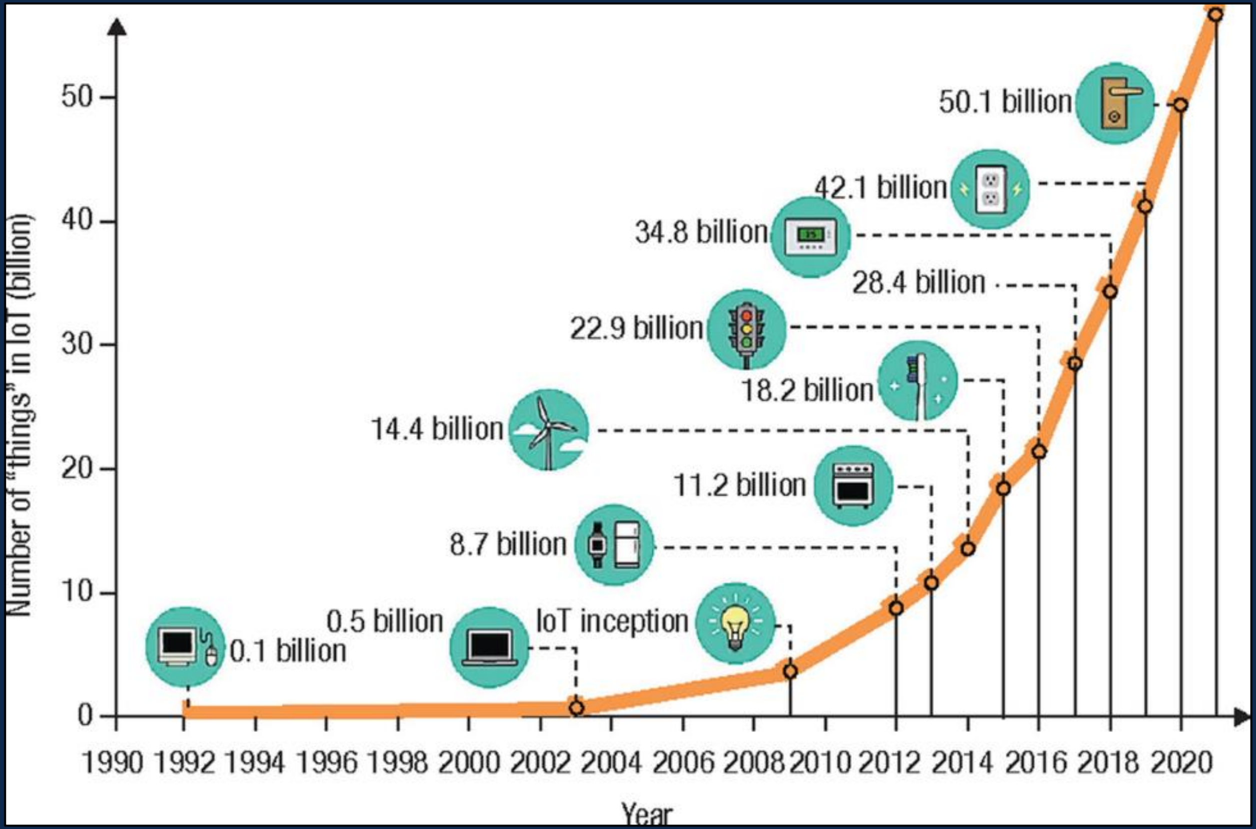
Jun 8, 2021

Data Today

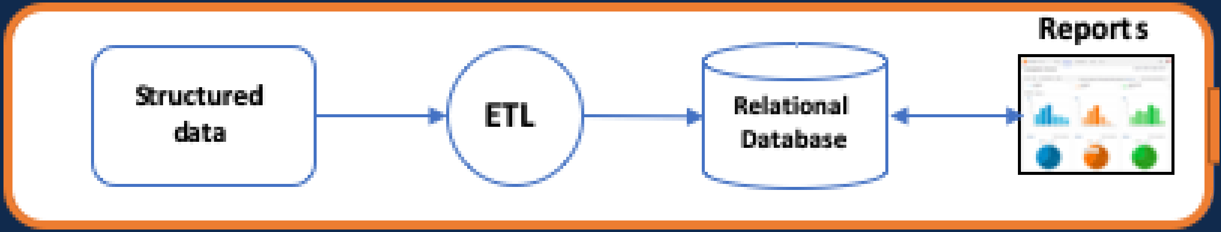


The growth of human and machine-generated data

Source: Inside big data

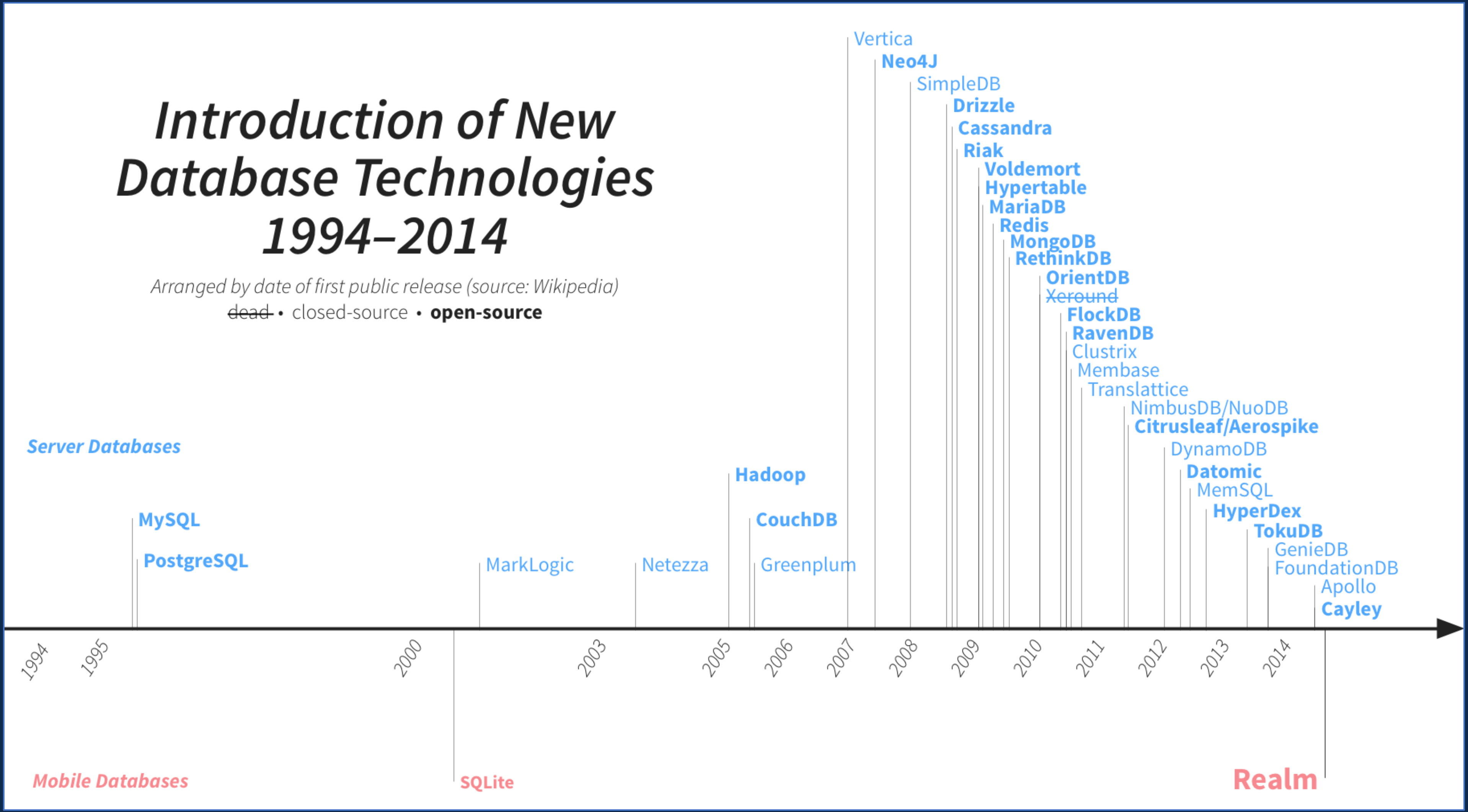


Traditional Data Warehouse



- Based on RDBMS (1980s)
- Meant for structured data
- Designed and build for a predefined purpose
- Purposefully rigid and not easily modified
- Data is cleaned and reformatted on upload (schema on write)
- Few users with advanced privileges to limit corruption and deletion
- Expensive to maintain, very difficult to scale
- *Can’t keep up with the volume, speed, granularity and demand of data today*

Race To Keep Up With Data Needs



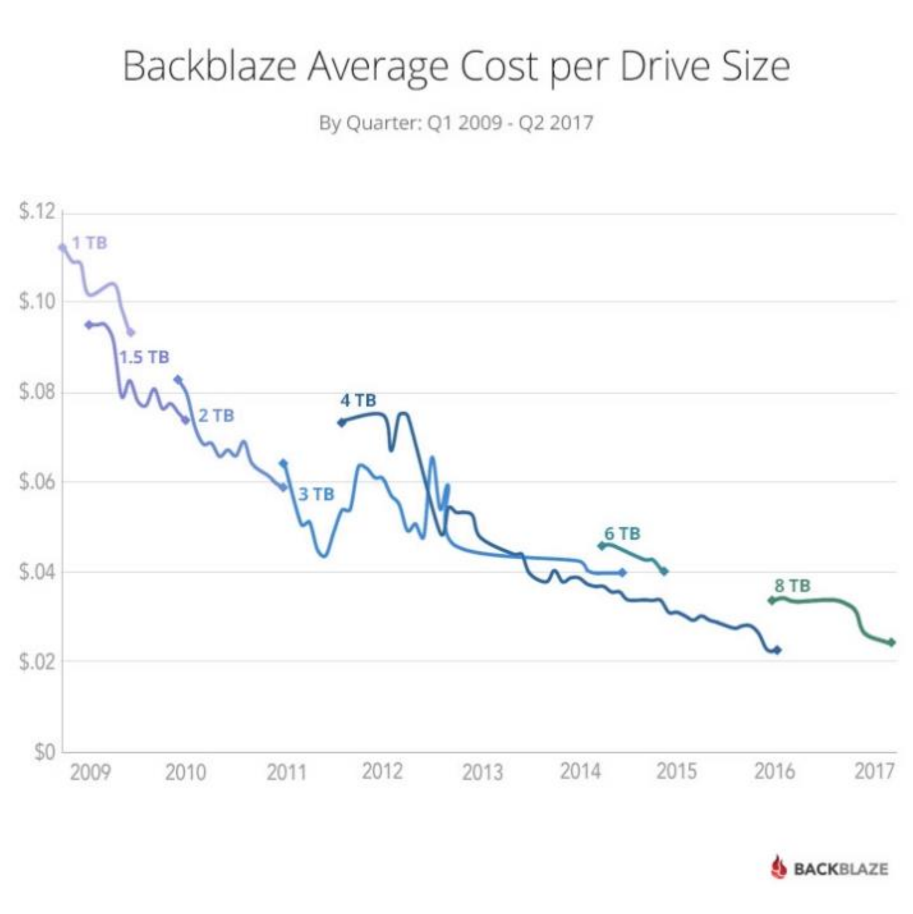
Introduction of New Database Technologies 1994–2014

Arranged by date of first public release (source: Wikipedia)

~~dead~~ • closed-source • **open-source**

Velocity of obsolescence

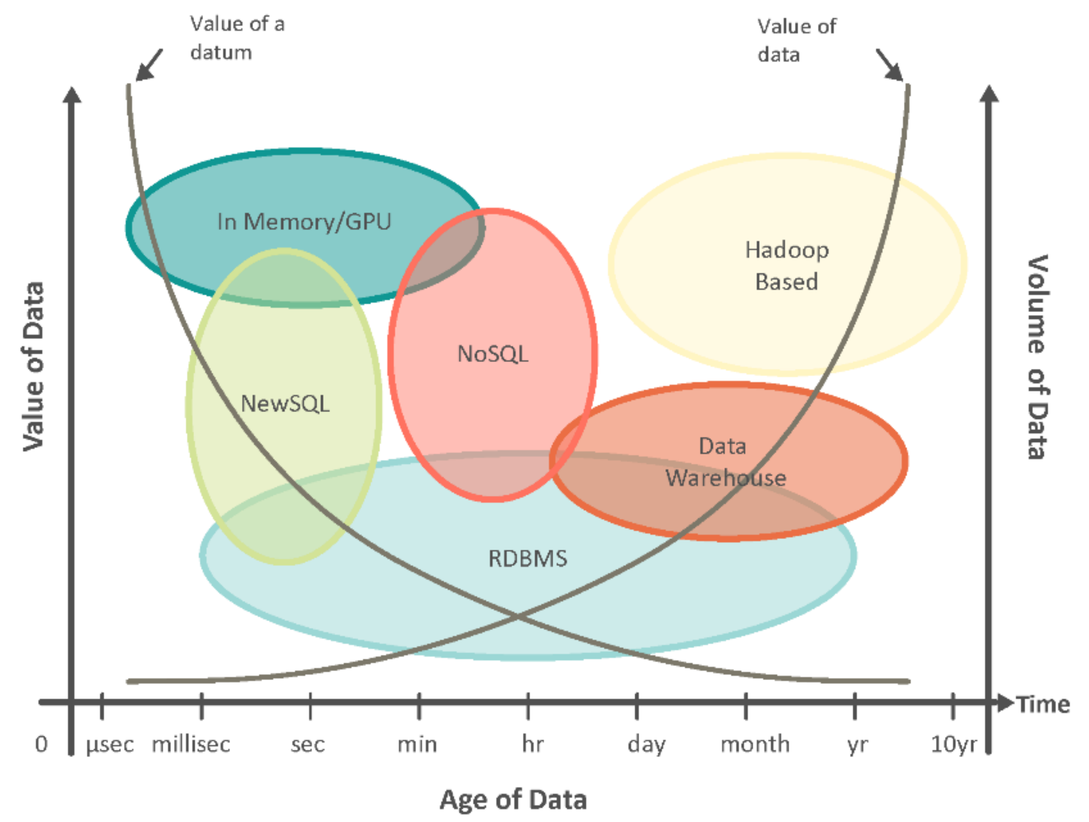
- **Obsolescence:** the time when a technical product or service is no longer needed or wanted even though it could still be in working order
- **Hardware:** Storage, Computational, Network
- **Software:** Automated, CI/CD
- **Workforce:** Half life of degrees and certifications is decreasing



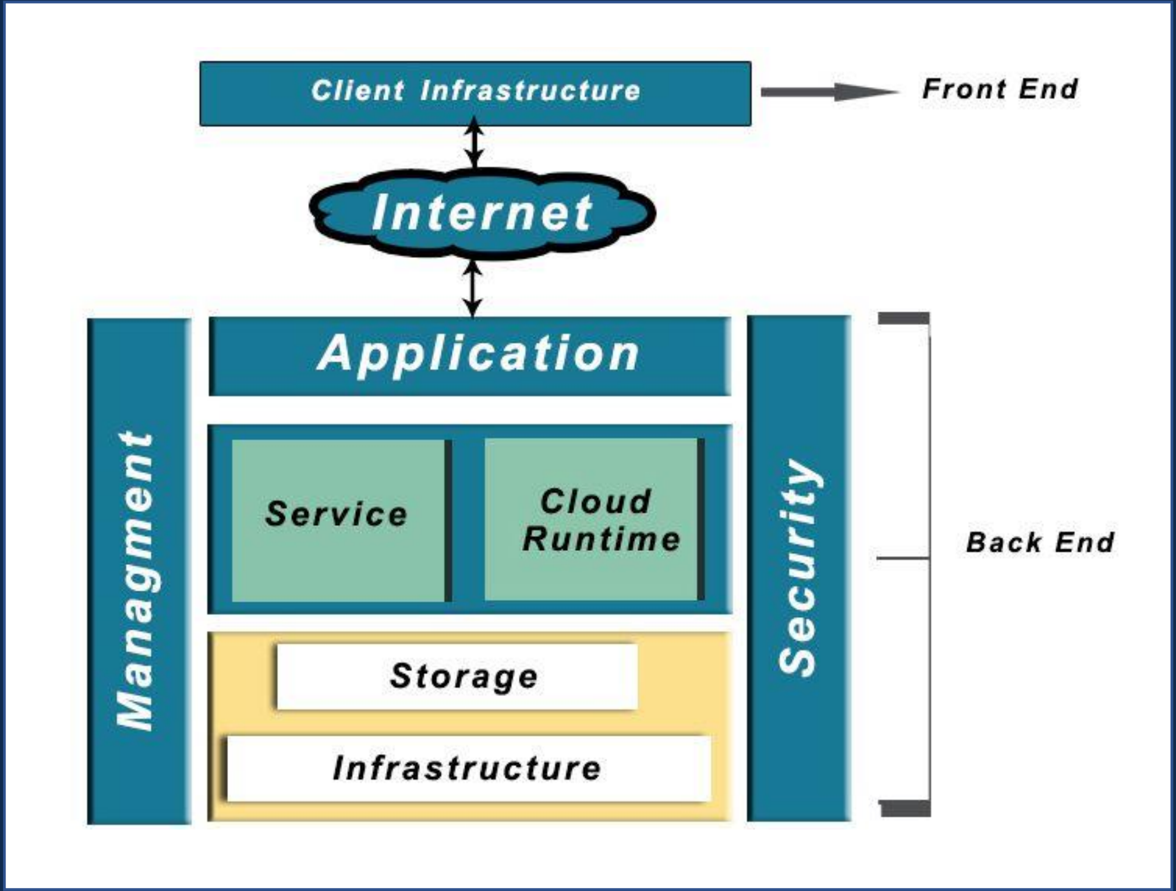
The Situation We Were In

- No more on size fits all solution
- Many data tools needed for equivalent RDBMS capabilities
- Tools need different hardware and networking
- Tools run on many servers (cluster)
- Excessive acquisition and maintenance costs

These shortcomings led to a new approach to shared IT resources - a.k.a. cloud

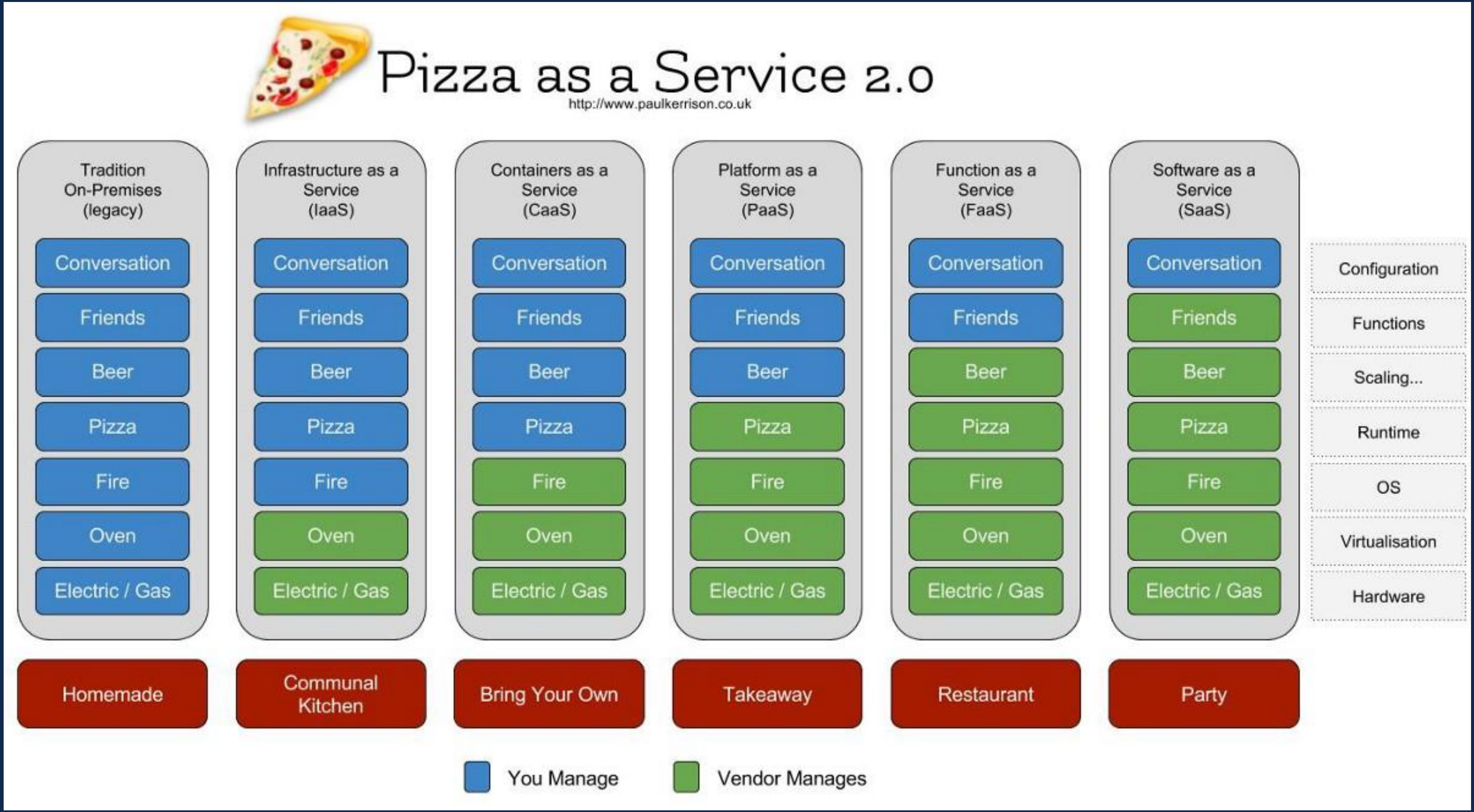


The Cloud



- On Demand
- Self Service
- Broad network access
- Multi-Tenancy (Resource Pooling)
- Rapid Elasticity
- Measured service
- Pay for what you use

Cloud Services



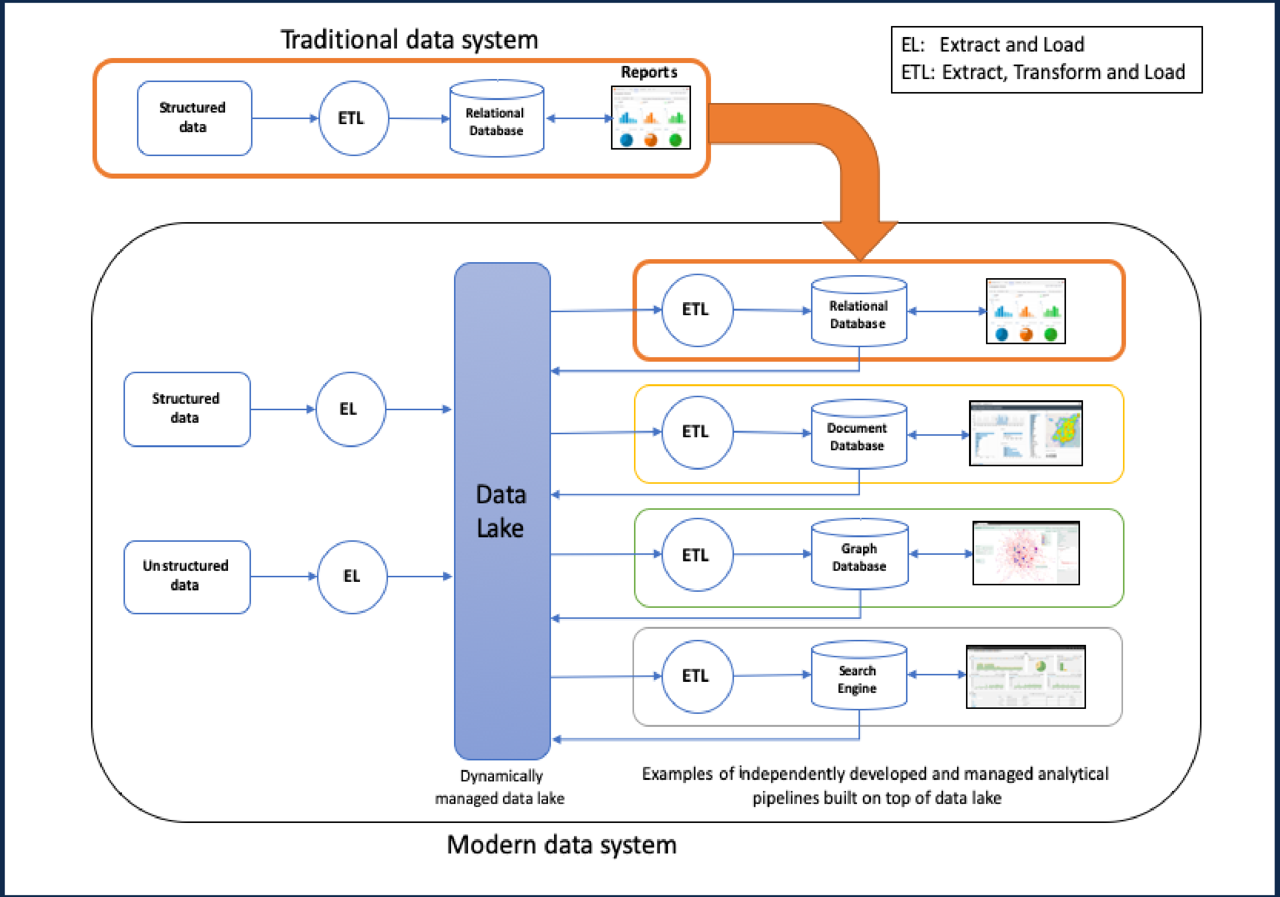
Pizza as a Service 2.0

http://www.paulkerrison.co.uk

Tradition On-Premises (legacy)	Infrastructure as a Service (IaaS)	Containers as a Service (CaaS)	Platform as a Service (PaaS)	Function as a Service (FaaS)	Software as a Service (SaaS)	
🟦 Conversation	🟦 Conversation	🟦 Conversation	🟦 Conversation	🟦 Conversation	🟦 Conversation	Configuration
🟦 Friends	🟦 Friends	🟦 Friends	🟦 Friends	🟦 Friends	🟩 Friends	Functions
🟦 Beer	🟦 Beer	🟦 Beer	🟦 Beer	🟩 Beer	🟩 Beer	Scaling...
🟦 Pizza	🟦 Pizza	🟦 Pizza	🟩 Pizza	🟩 Pizza	🟩 Pizza	Runtime
🟦 Fire	🟦 Fire	🟩 Fire	🟩 Fire	🟩 Fire	🟩 Fire	OS
🟦 Oven	🟩 Oven	🟩 Oven	🟩 Oven	🟩 Oven	🟩 Oven	Virtualisation
🟦 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	🟩 Electric / Gas	Hardware
Homemade	Communal Kitchen	Bring Your Own	Takeaway	Restaurant	Party	

🟦 You Manage 🟩 Vendor Manages

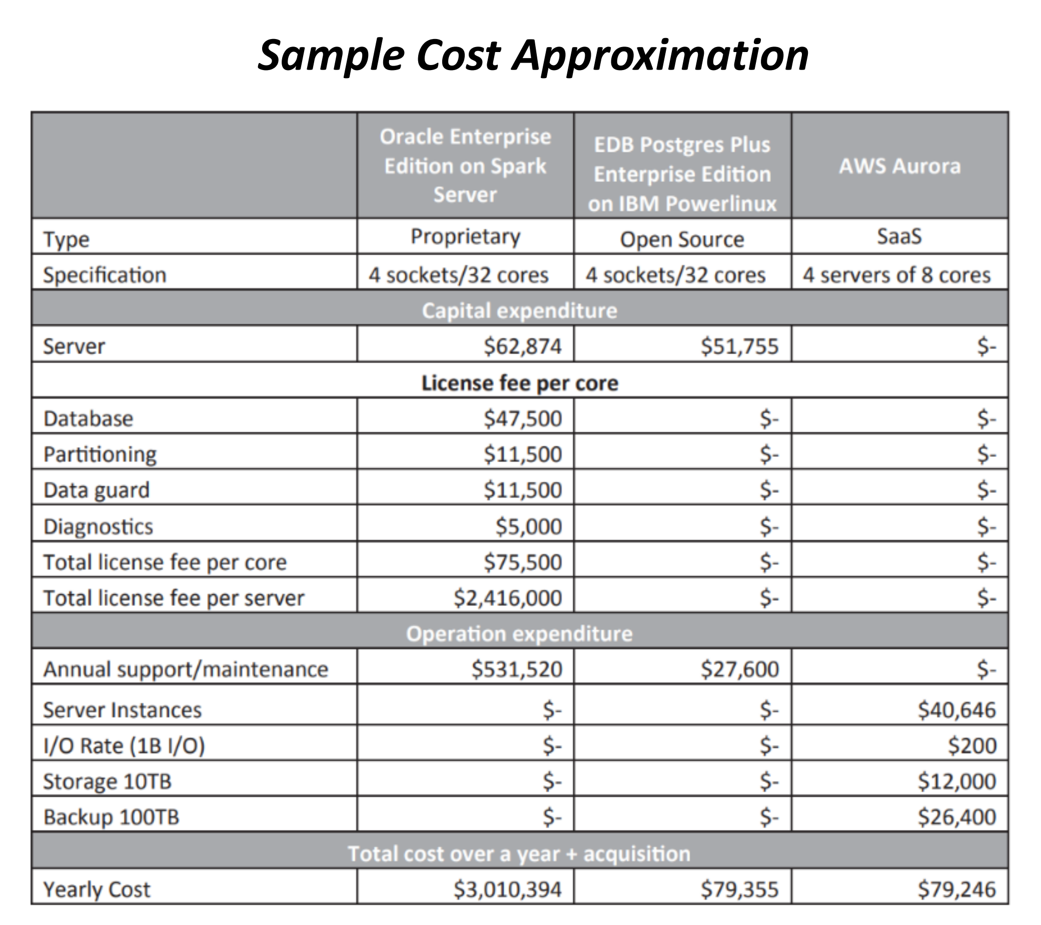
The Modern Data Warehouse



- Load data as is, no cleaning or reformatting
- Scalable storage (data lake)
- Extract, transform, load and analyze data differently for each use case
- The traditional data warehouse using RDBMS still works with modern framework.

Cost Comparison

- **Less expensive, but based on demand**
- **Need to be vigilant of change**
- **Watch out for demand surge**
  - Snowstorm
  - Long running queries
  - Disconnected software instances (Ghost)
  - Forgotten data



Sample Cost Approximation

	Oracle Enterprise Edition on Spark Server	EDB Postgres Plus Enterprise Edition on IBM Powerlinux	AWS Aurora
Type	Proprietary	Open Source	SaaS
Specification	4 sockets/32 cores	4 sockets/32 cores	4 servers of 8 cores
Capital expenditure			
Server	$62,874	$51,755	$-
License fee per core			
Database	$47,500	$-	$-
Partitioning	$11,500	$-	$-
Data guard	$11,500	$-	$-
Diagnostics	$5,000	$-	$-
Total license fee per core	$75,500	$-	$-
Total license fee per server	$2,416,000	$-	$-
Operation expenditure			
Annual support/maintenance	$531,520	$27,600	$-
Server Instances	$-	$-	$40,646
I/O Rate (1B I/O)	$-	$-	$200
Storage 10TB	$-	$-	$12,000
Backup 100TB	$-	$-	$26,400
Total cost over a year + acquisition			
Yearly Cost	$3,010,394	$79,355	$79,246

Not So Fast

Why On-Premise

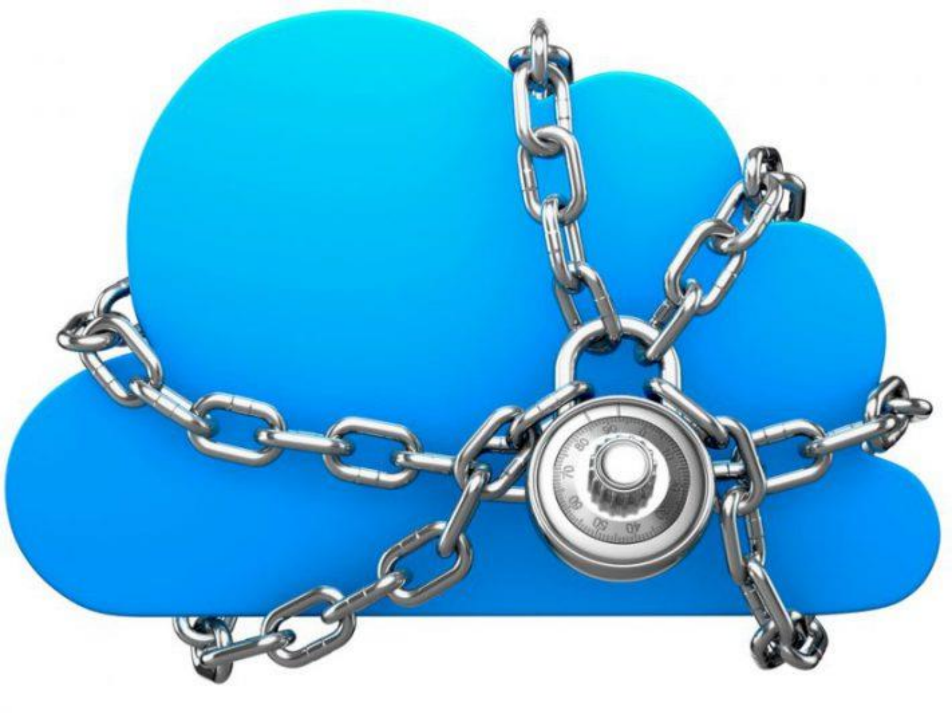
- To meet government regulations for storing sensitive data
- Need for unique/advanced security beyond cloud offering
- Visibility of data “residency”
- Bandwidth constraining accessibility at the last mile
- More direct control over latency

Why Cloud

- Shifts the risk of IT infrastructure obsolescence to the cloud provider
- Enables a scalable, flexible and on demand set of IT capabilities
- Reduce IT infrastructure operation and maintenance time
- Lower cost profile

What's the catch – 1 of 2

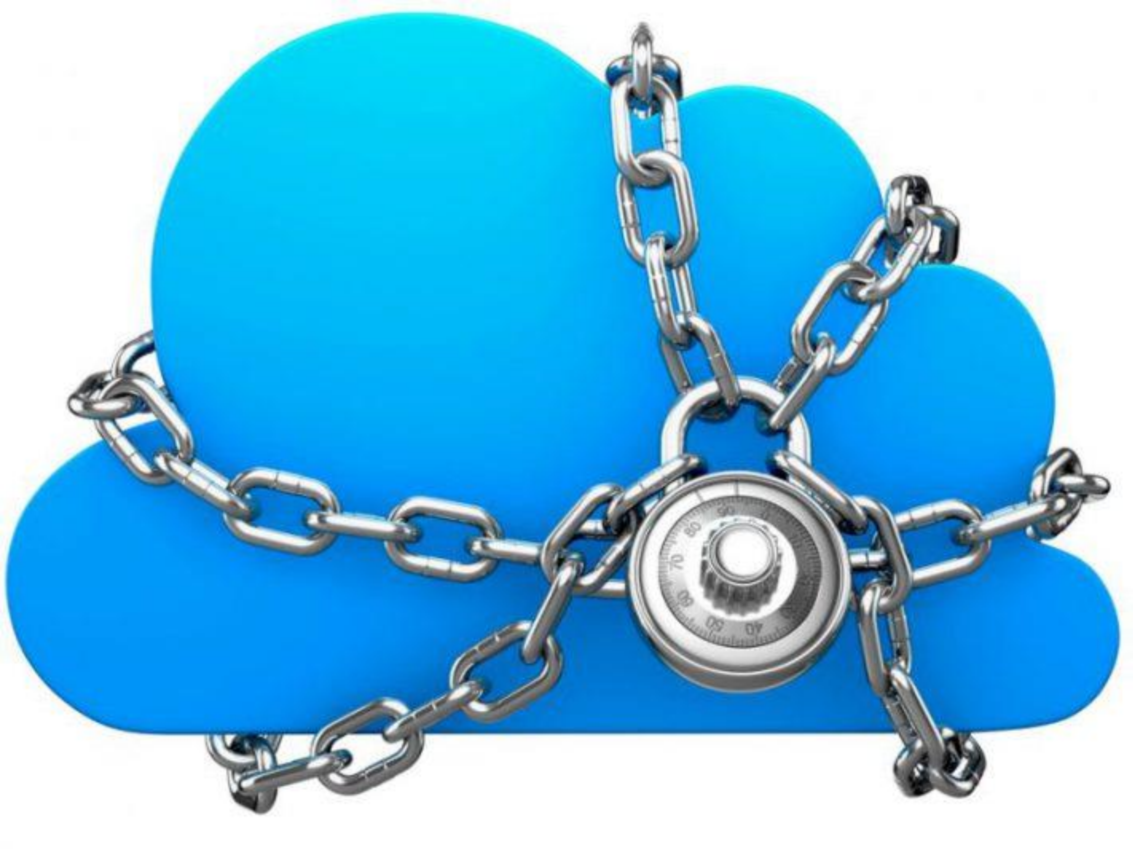
- If you transfer your current architecture as is…
  - There will be minimal to no real benefits
  - Potentially increase costs compare to on premise
- You need to rearchitect to take advantage of cloud services
  - To be scalable
  - To be resilient
  - To pay only for what you use



What's the catch – 2 of 2

- Rearchitecting will include, at a minimum, focusing on:
  - Deploying software differently
  - Managing data differently

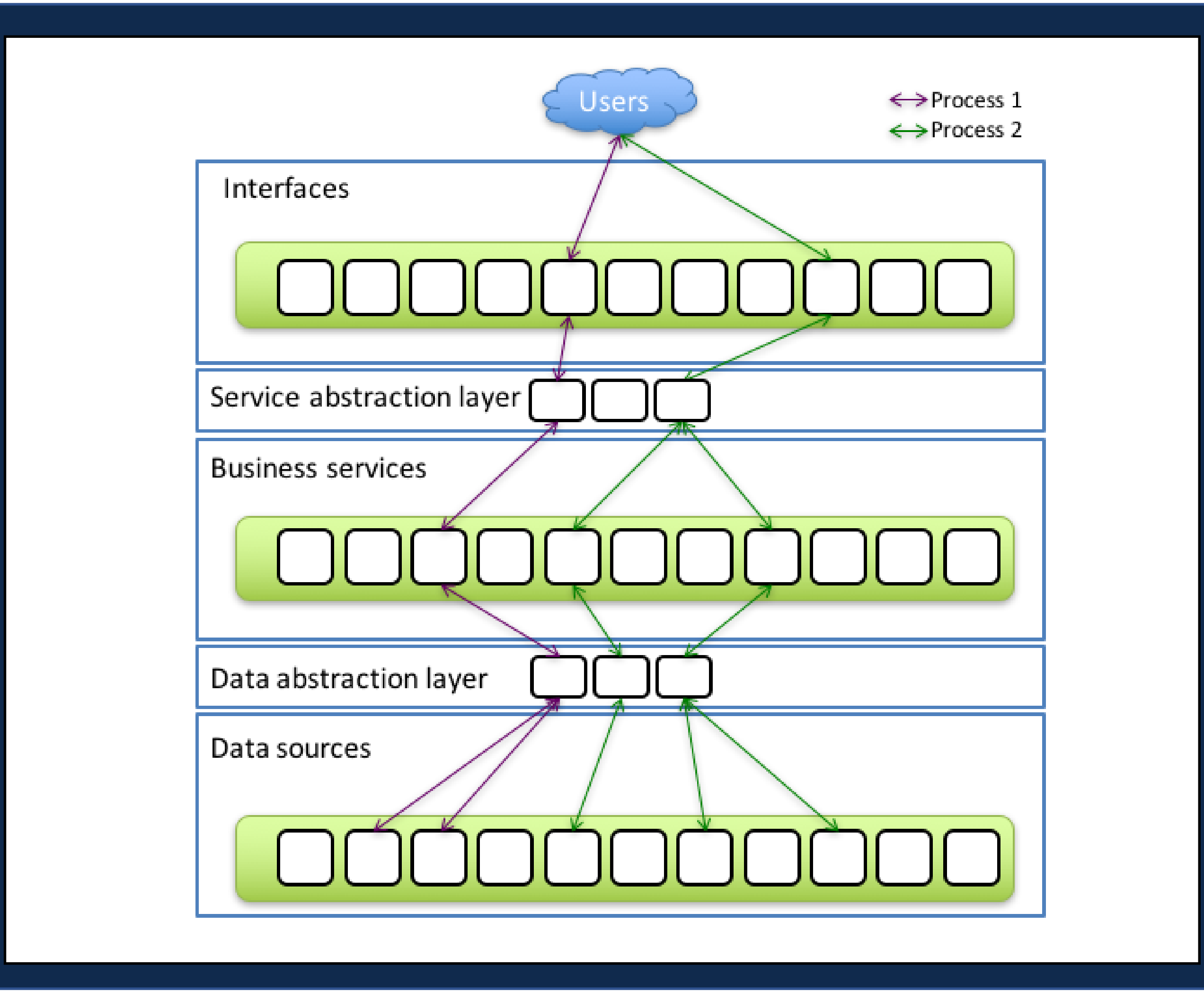
Beware of vendor lock!



Distributed software architecture – 1 of 2

Distributed system: an IT system in which the computing power and software is:

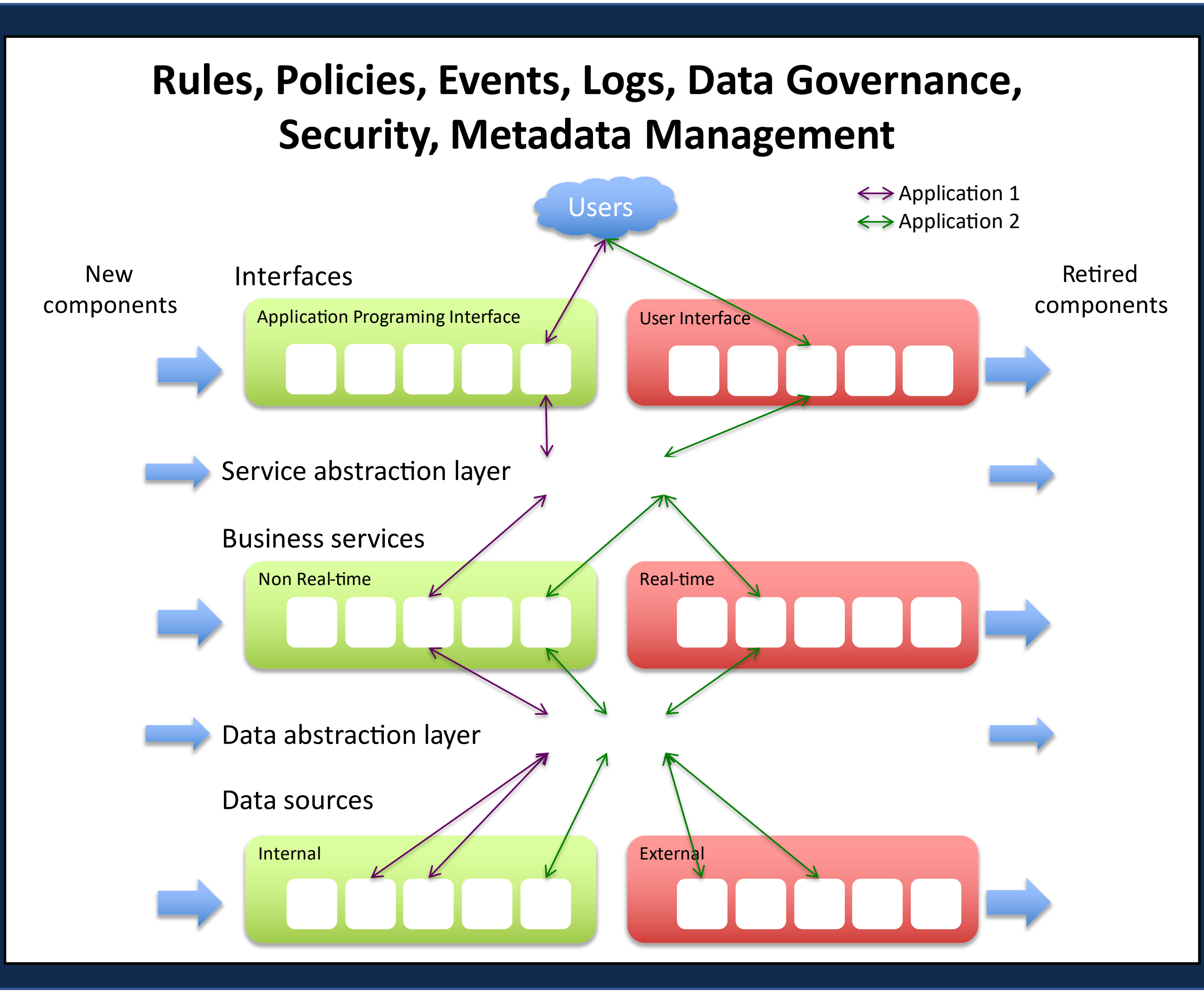
- Distributed across several servers,
- Connected through a network, communicating, and
- Coordinating their actions by passing messages to each other.



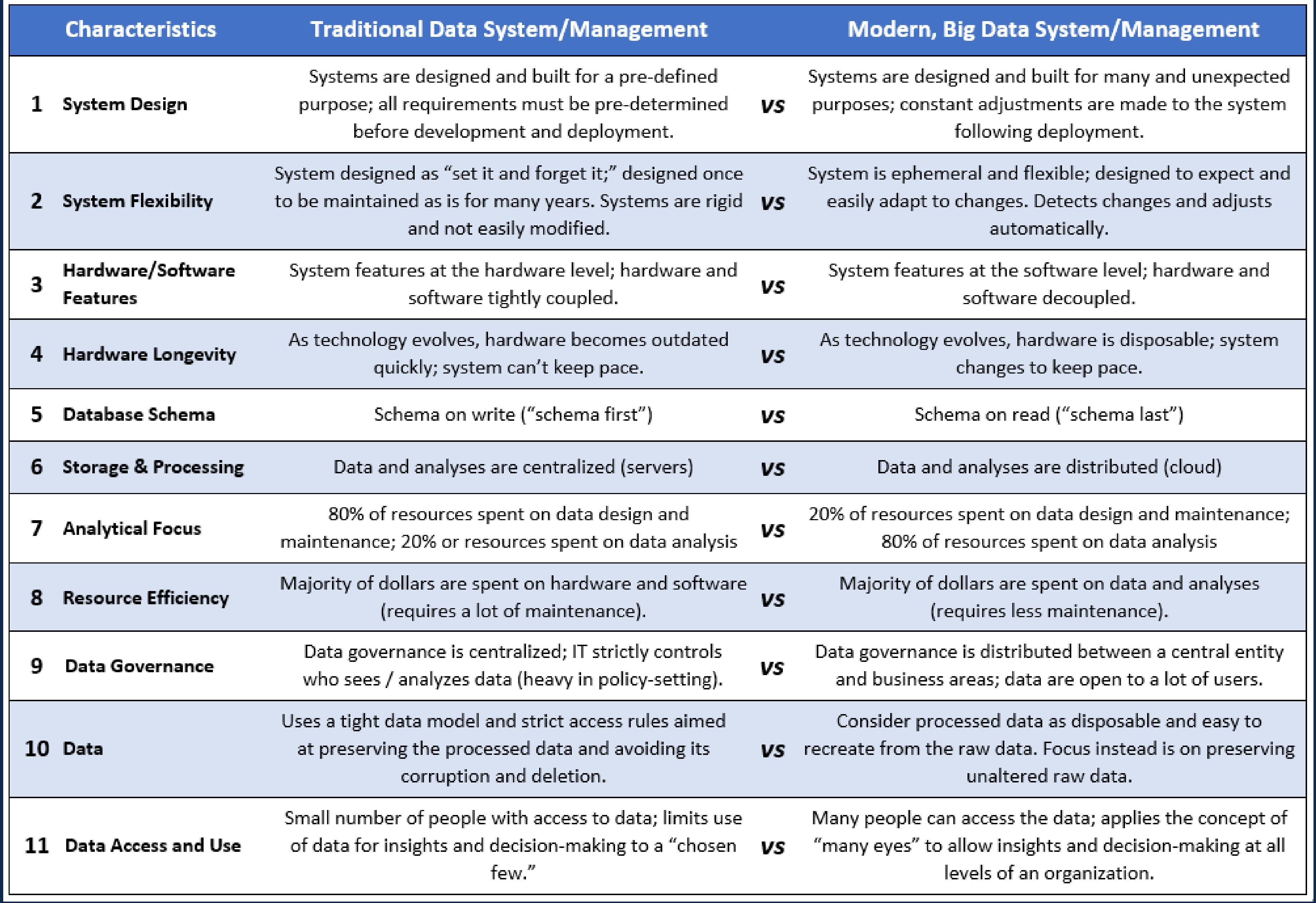
Distributed software architecture – 2 of 2

The Distributed System requires attention to:

- Rules and Policies
- Events and Logs
- Data Governance
- Security
- Metadata Management

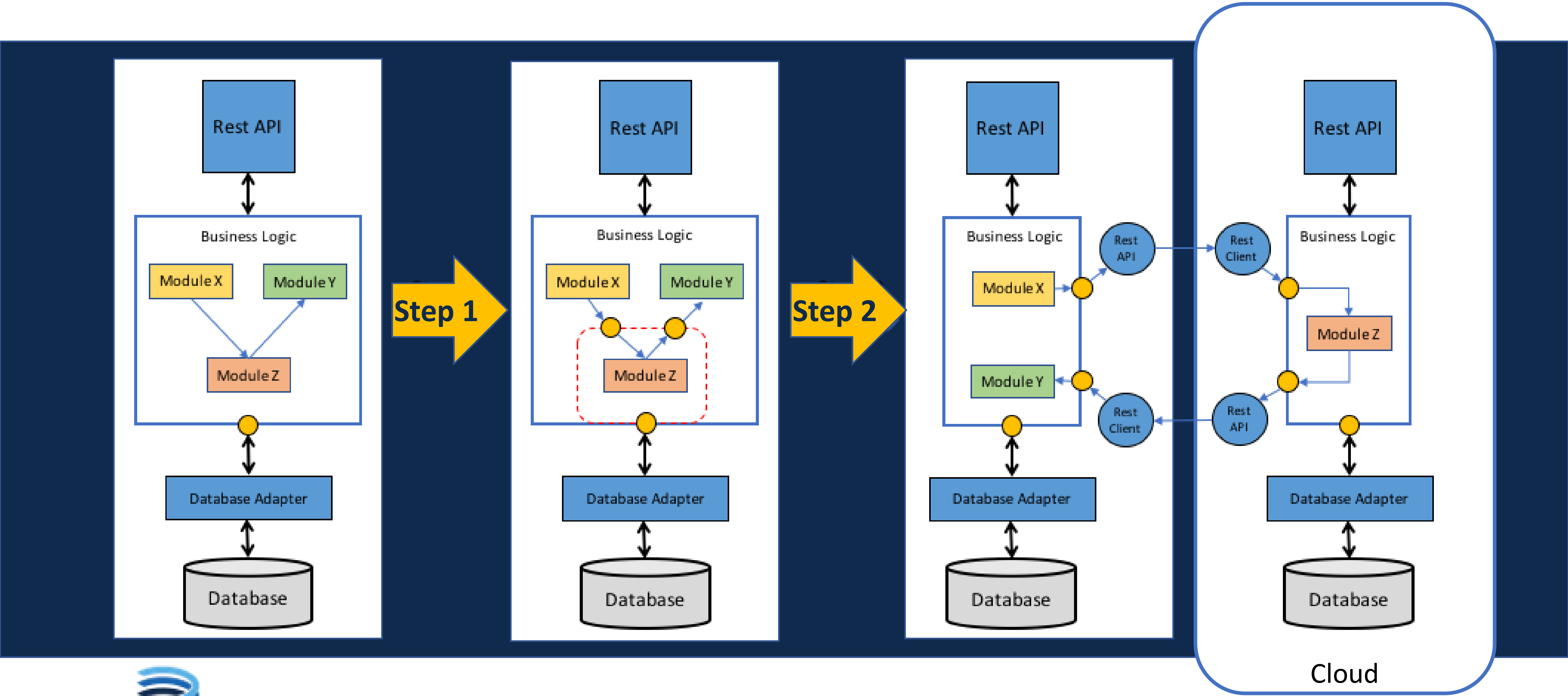


Traditional v. Modern Data Management



	Characteristics	Traditional Data System/Management		Modern, Big Data System/Management
1	System Design	Systems are designed and built for a pre-defined purpose; all requirements must be pre-determined before development and deployment.	VS	Systems are designed and built for many and unexpected purposes; constant adjustments are made to the system following deployment.
2	System Flexibility	System designed as “set it and forget it;” designed once to be maintained as is for many years. Systems are rigid and not easily modified.	VS	System is ephemeral and flexible; designed to expect and easily adapt to changes. Detects changes and adjusts automatically.
3	Hardware/Software Features	System features at the hardware level; hardware and software tightly coupled.	VS	System features at the software level; hardware and software decoupled.
4	Hardware Longevity	As technology evolves, hardware becomes outdated quickly; system can’t keep pace.	VS	As technology evolves, hardware is disposable; system changes to keep pace.
5	Database Schema	Schema on write (“schema first”)	VS	Schema on read (“schema last”)
6	Storage & Processing	Data and analyses are centralized (servers)	VS	Data and analyses are distributed (cloud)
7	Analytical Focus	80% of resources spent on data design and maintenance; 20% or resources spent on data analysis	VS	20% of resources spent on data design and maintenance; 80% of resources spent on data analysis
8	Resource Efficiency	Majority of dollars are spent on hardware and software (requires a lot of maintenance).	VS	Majority of dollars are spent on data and analyses (requires less maintenance).
9	Data Governance	Data governance is centralized; IT strictly controls who sees / analyzes data (heavy in policy-setting).	VS	Data governance is distributed between a central entity and business areas; data are open to a lot of users.
10	Data	Uses a tight data model and strict access rules aimed at preserving the processed data and avoiding its corruption and deletion.	VS	Consider processed data as disposable and easy to recreate from the raw data. Focus instead is on preserving unaltered raw data.
11	Data Access and Use	Small number of people with access to data; limits use of data for insights and decision-making to a “chosen few.”	VS	Many people can access the data; applies the concept of “many eyes” to allow insights and decision-making at all levels of an organization.

How To Transition? Slowly But Surely



Want to know more?

- **NCHRP Research Report 865**

  Guide for Development and Management of Sustainable Enterprise Information Portals

- **NCHRP Research Report 952**

  Guidebook for Managing Data from Emerging Technologies for Transportation

- **NCHRP Research report 904**

  Leveraging Big Data to Improve Traffic Incident Management

Benjamin Pecheux
Director of Information Research
AEM Corporation and FWHA
Crowdsourcing Contract Support
Benjamin.pecheux@aemcorp.com