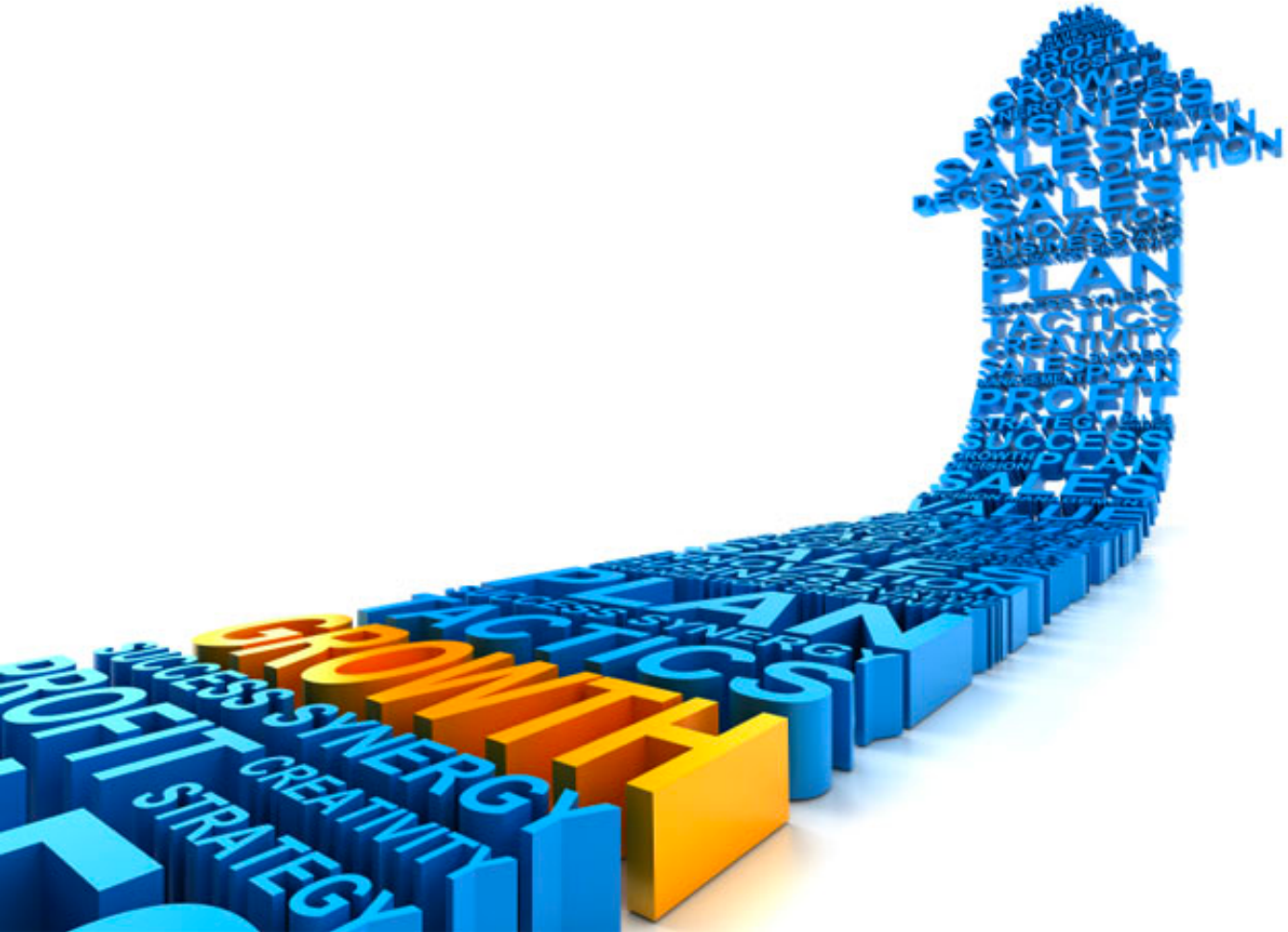
IT University of Copenhagen

Big Data Management: Lambda Architecture

Björn Þór Jónsson

Big Data
Building Systems

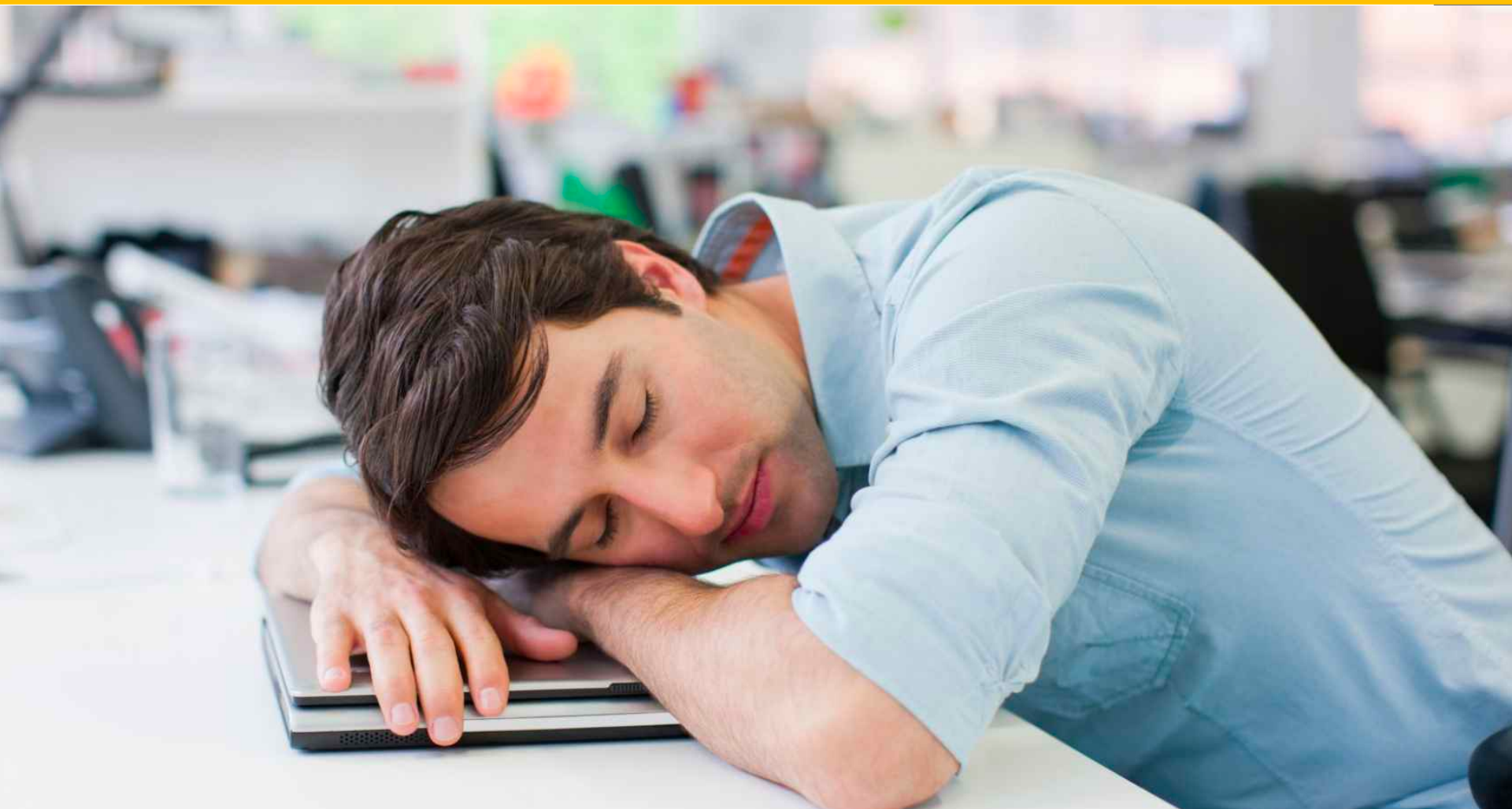
Requirements
Scalability



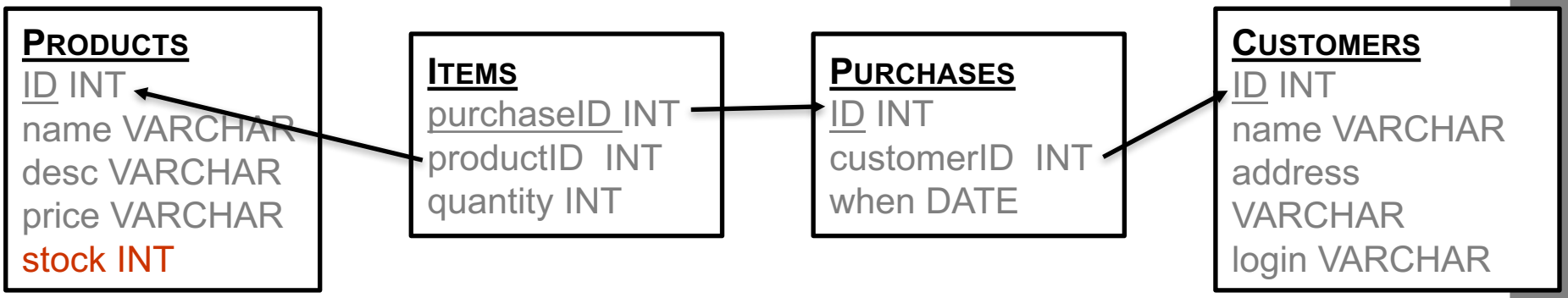
Requirements
Low Latency

Requirements
Fault Tolerance

Requirements
Easy Maintenance

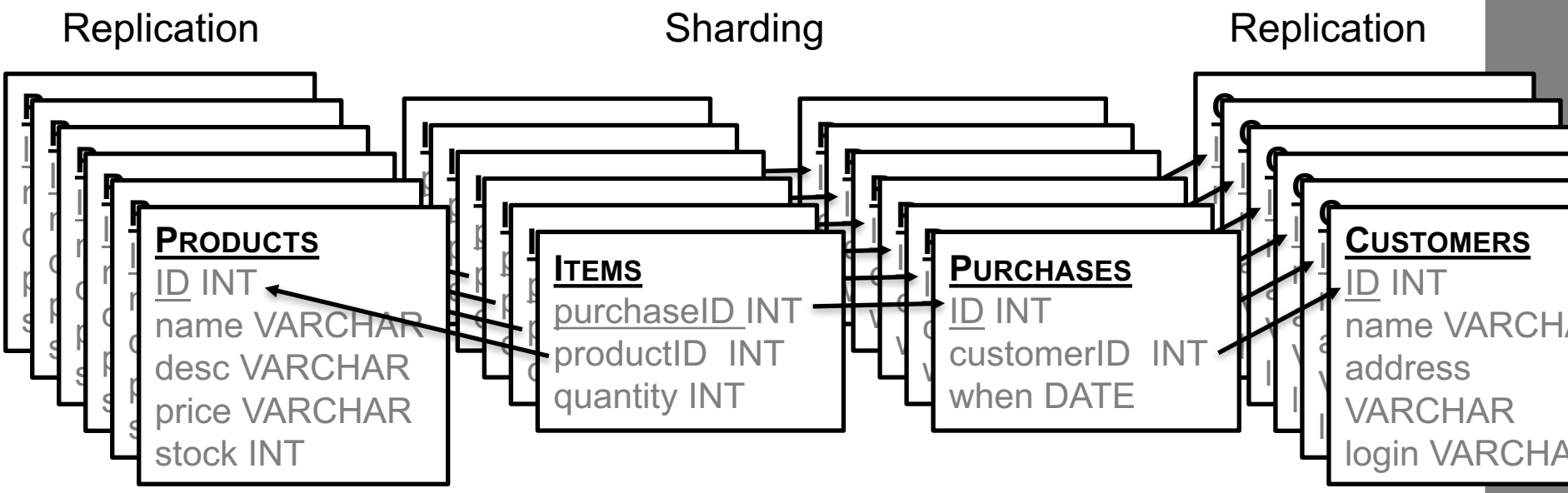


Traditional Solution vs Requirements

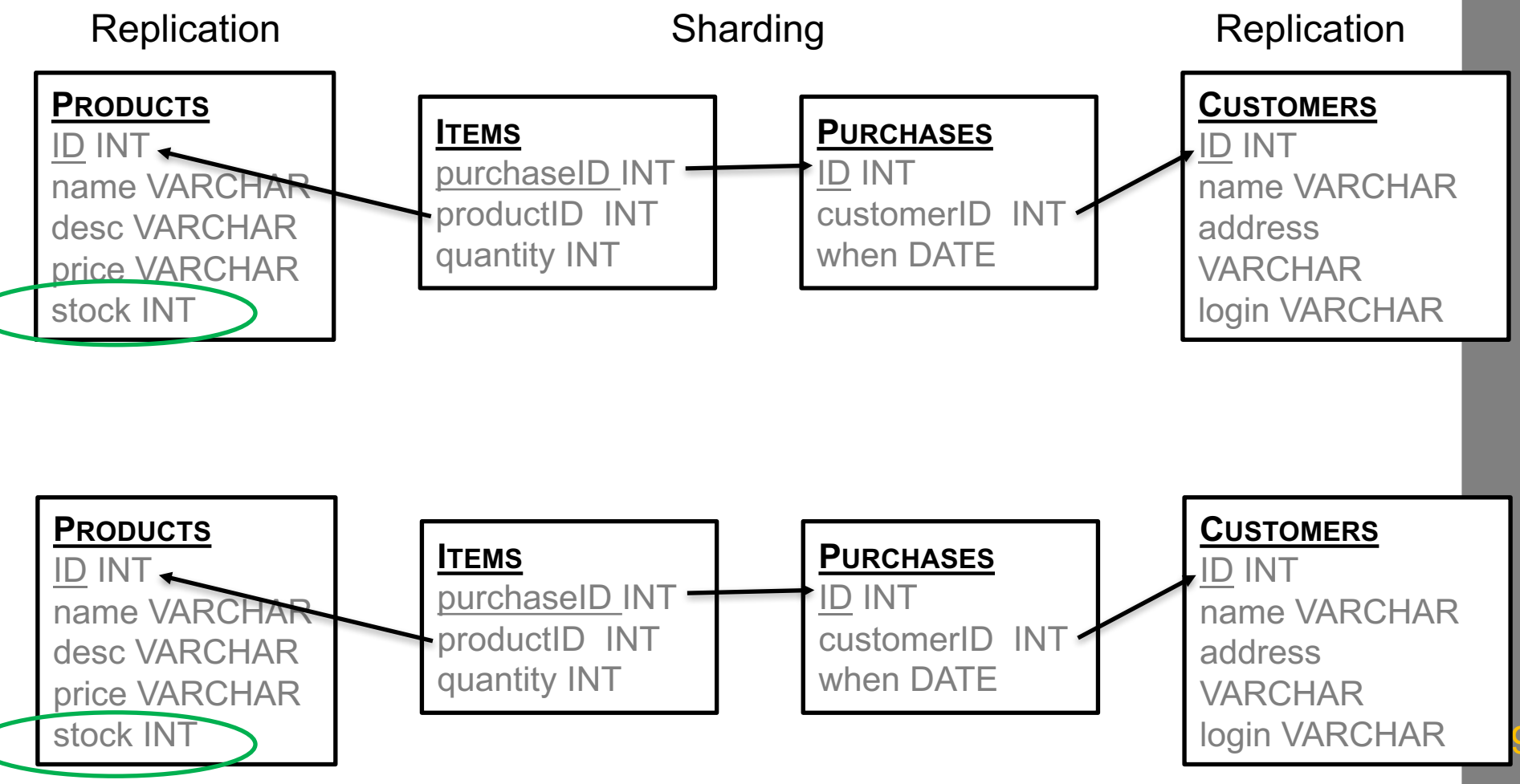


- Scalability?
- Low latency?
- Fault tolerance?
- Maintainability?

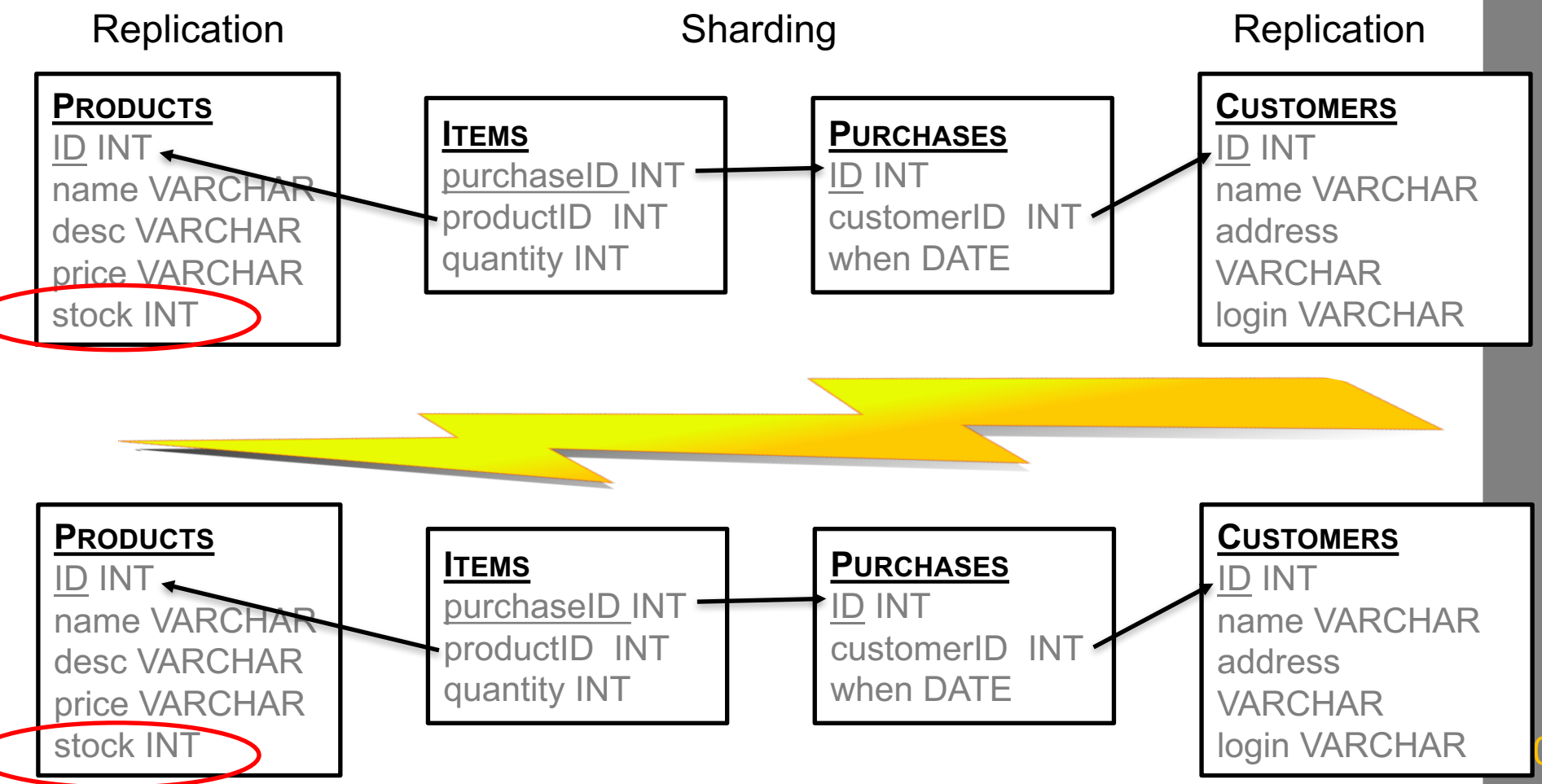
Scaling Traditional Solutions



Scaling Traditional Solutions



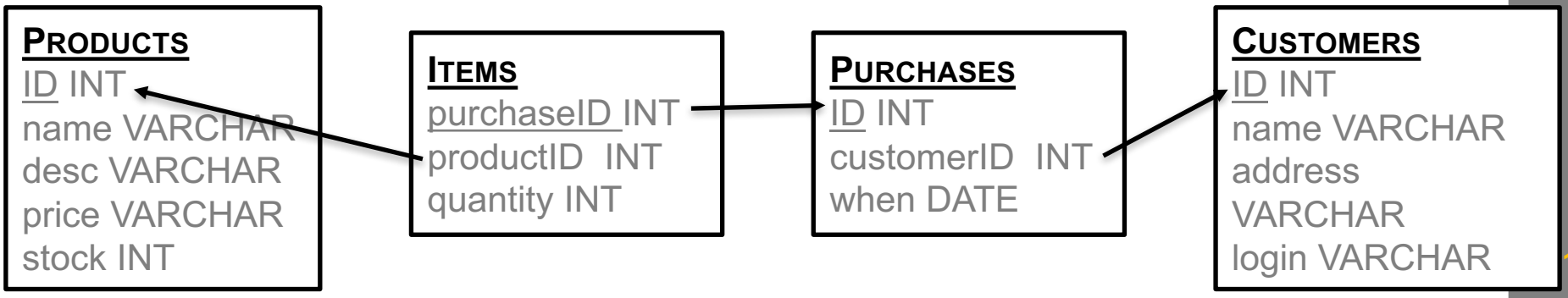
Scaling Traditional Solutions



Fault Tolerance Traditional Solutions

- What happens with human errors?
  - New transaction does not update stock…
  - Administrator deletes part of Items table...

IT University of Copenhagen



Traditional Solution vs Requirements

- Scalability?
- Low latency?
- Fault tolerance?
- Maintainability?

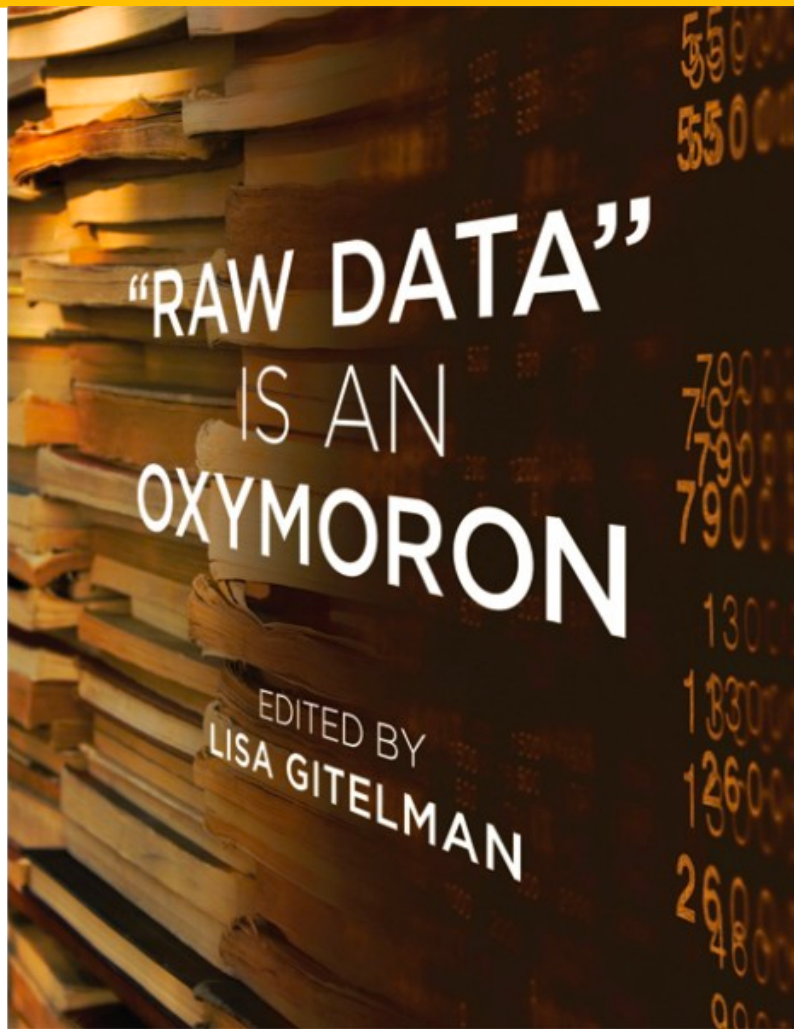
Conflicting Requirements!

Data Properties

Data is raw

Data is immutable

Data is forever



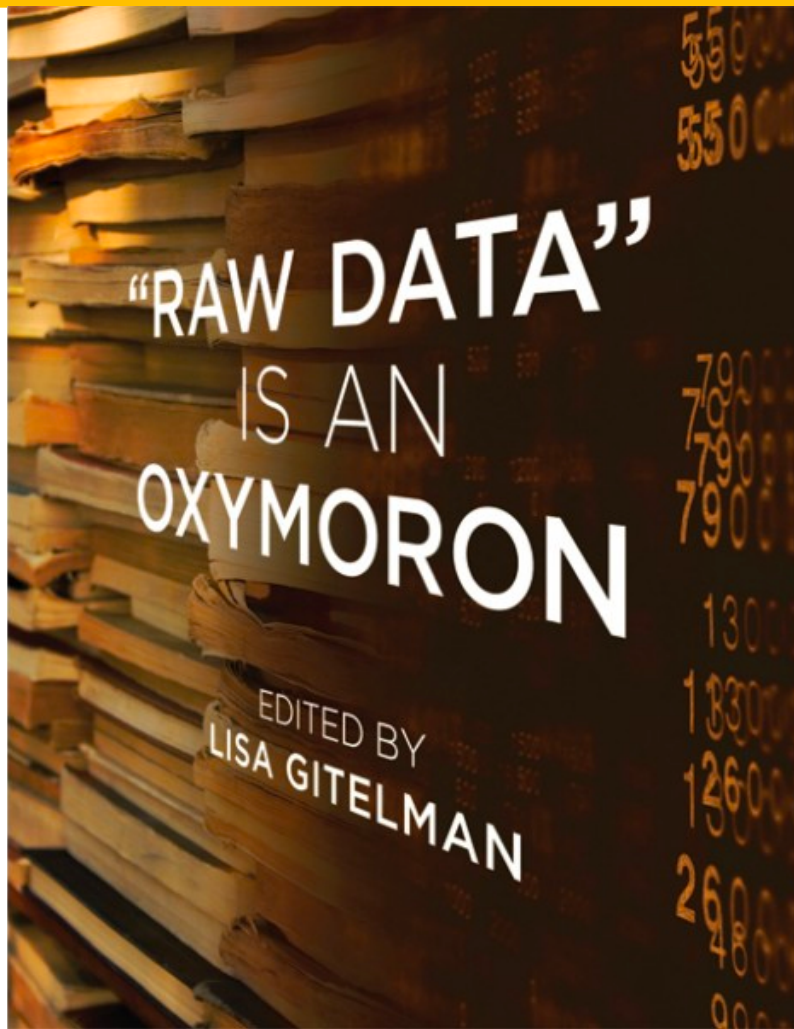
Software Engineers Guardians of the Data

Data Properties

Data is raw

Data is immutable

Data is forever

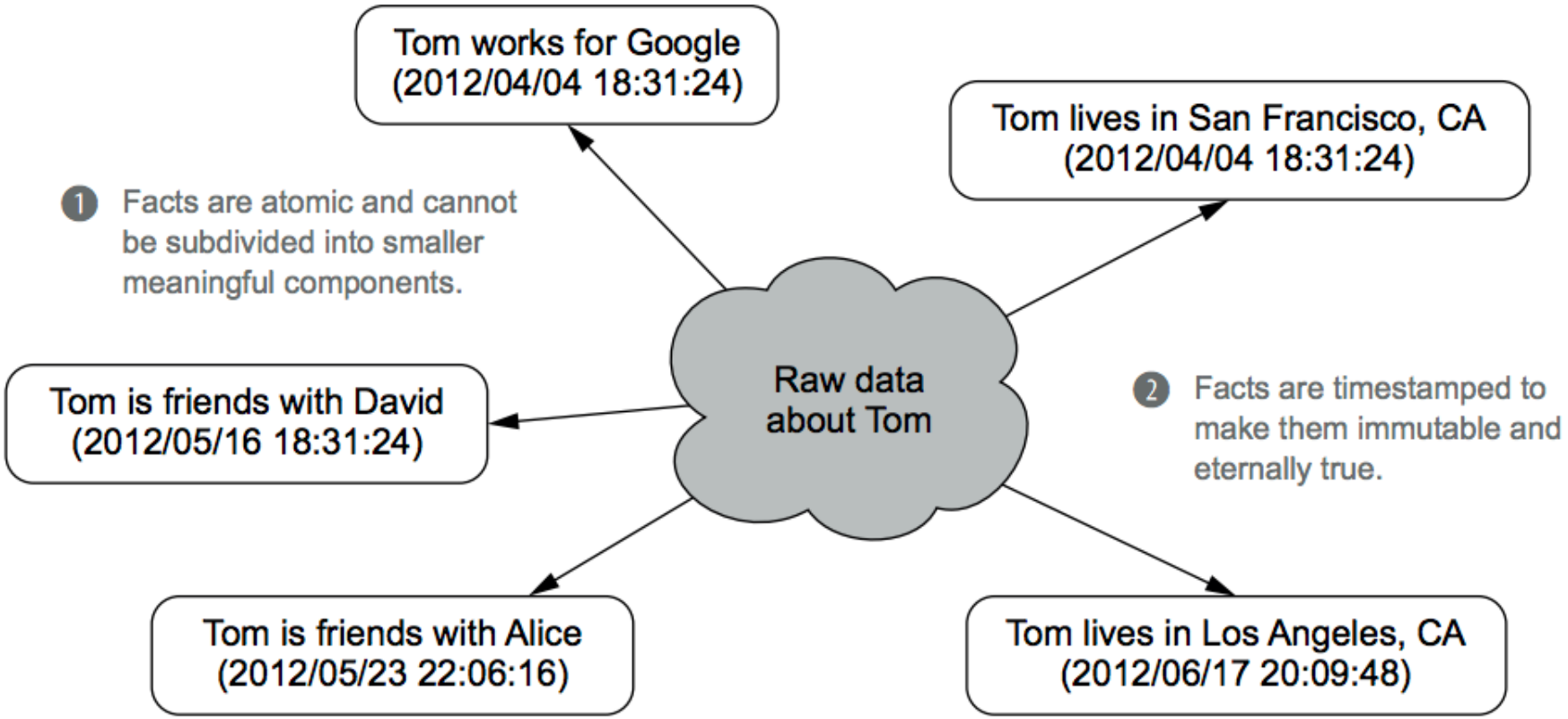


Security and Data Retention

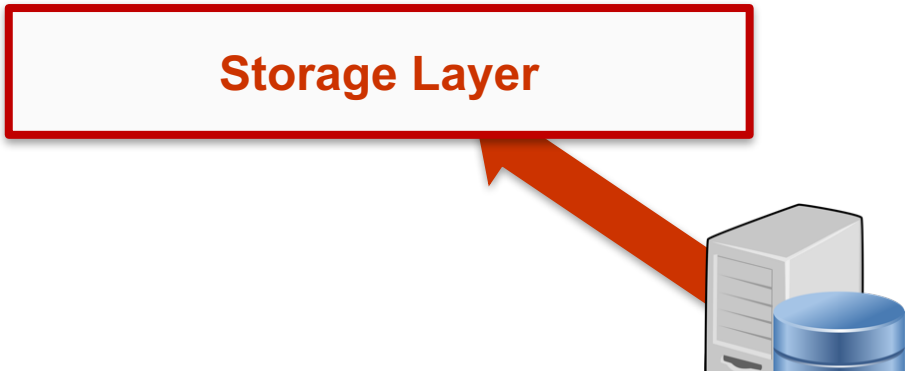
And of course, if you don't need to keep data, don't store it. Having healthy retention policies is important. Adkins also uses the same strategy for her own data. "I delete all the love letters from my husband," she said.

https://techcrunch.com/2017/09/18/googles-heather-adkins-thinks-everybody-is-going-to-get-hacked-and-you-need-to-be-ready/

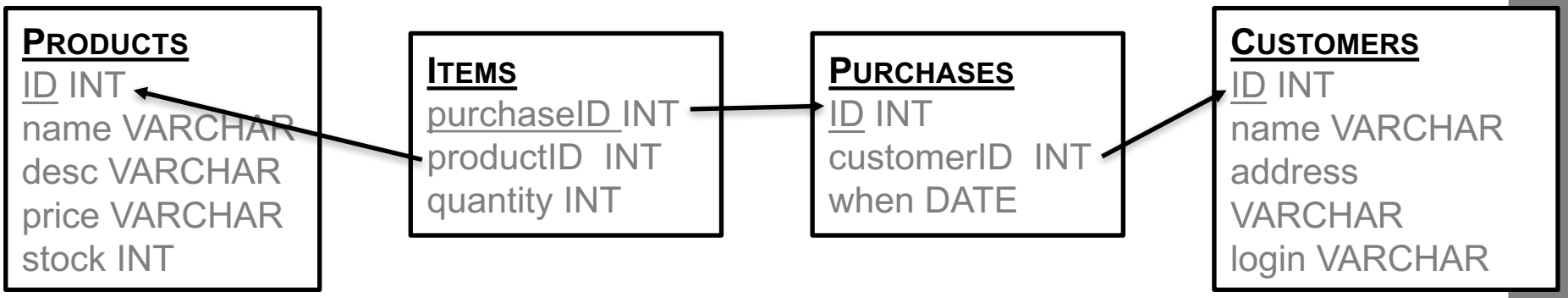
Fact-Based Data Model



Big Data Framework: Lambda Architecture



Low Latency ~ Current State

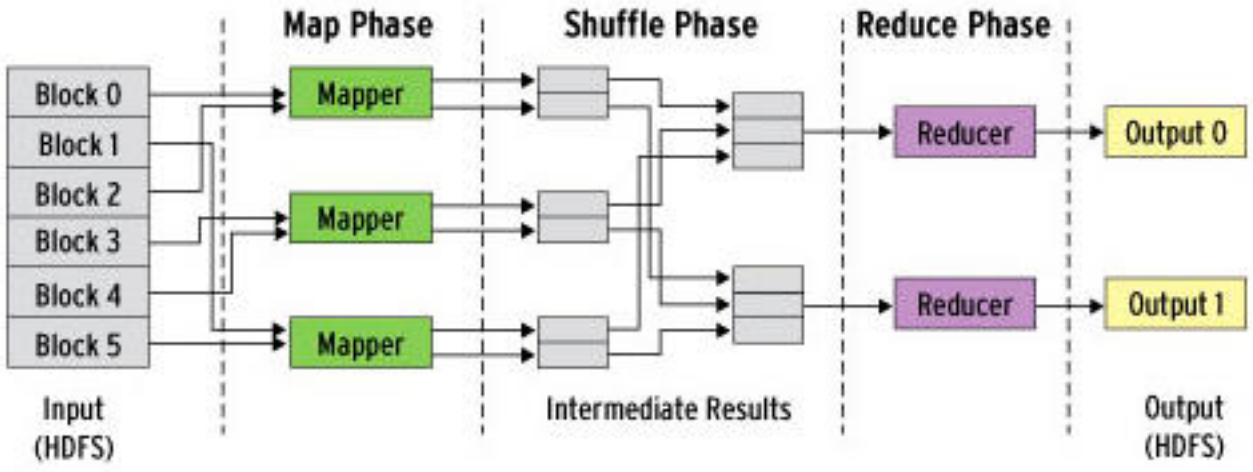


Big Data Framework: Lambda Architecture

Service Layer

Storage Layer

Storage Layer to Service Layer



APACHE Spark™

Big Data Framework: Lambda Architecture

Service Layer

Batch Layer

Storage Layer

Latency

Throughput

Volume

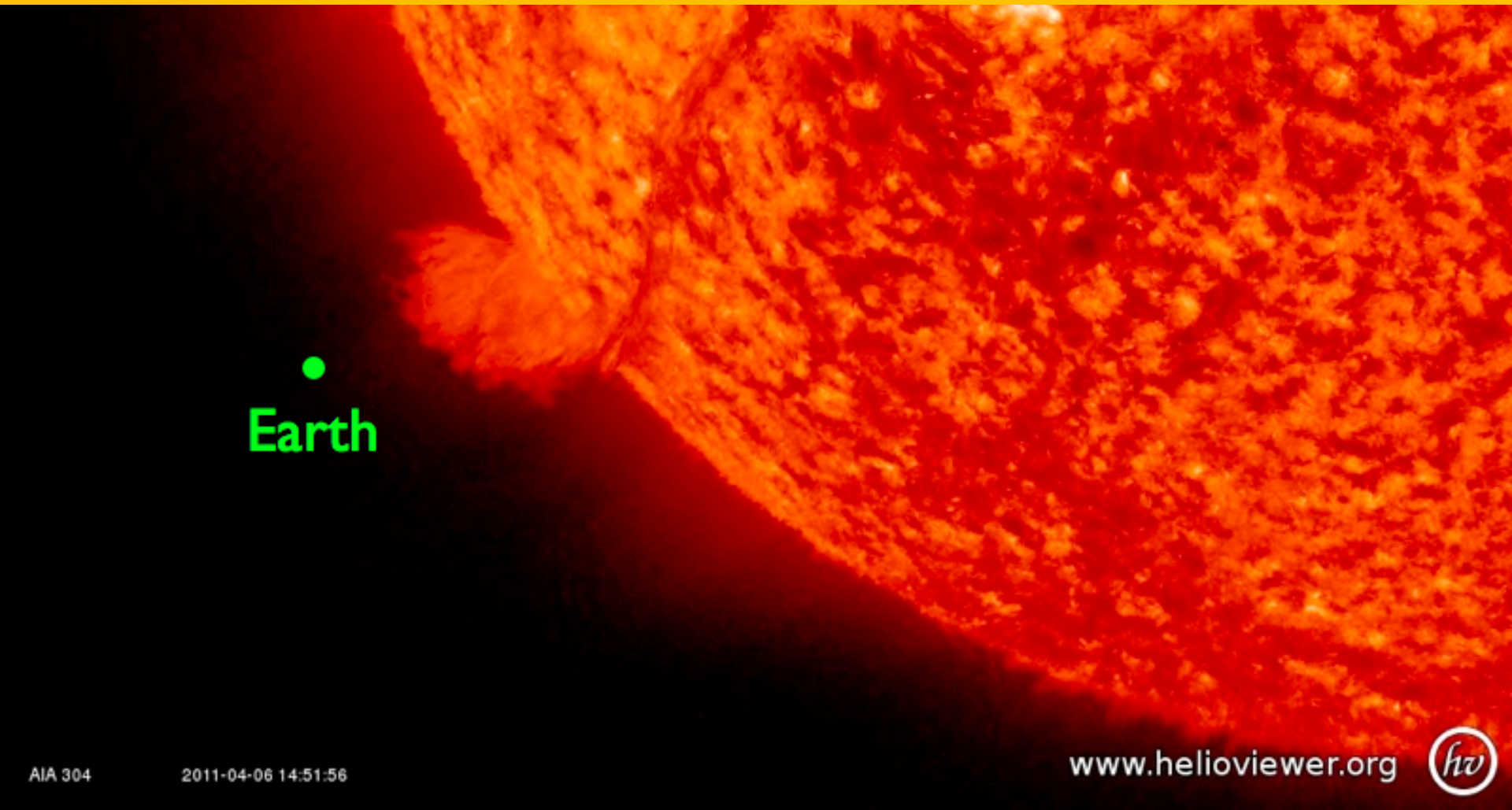
Batch Processing Takes Time

Service Layer

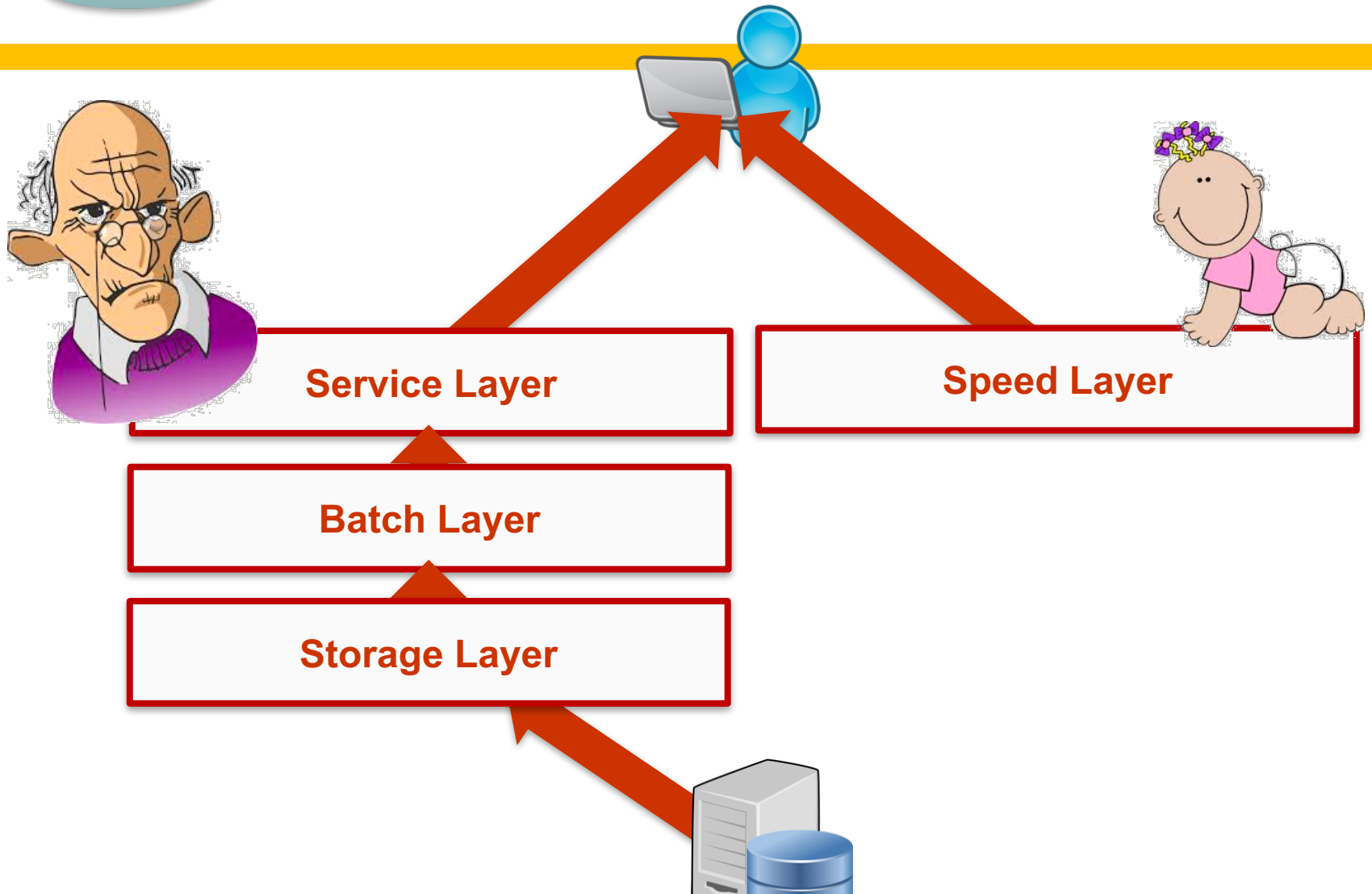
Batch Layer

Storage Layer

Stale Data vs Fresh Data



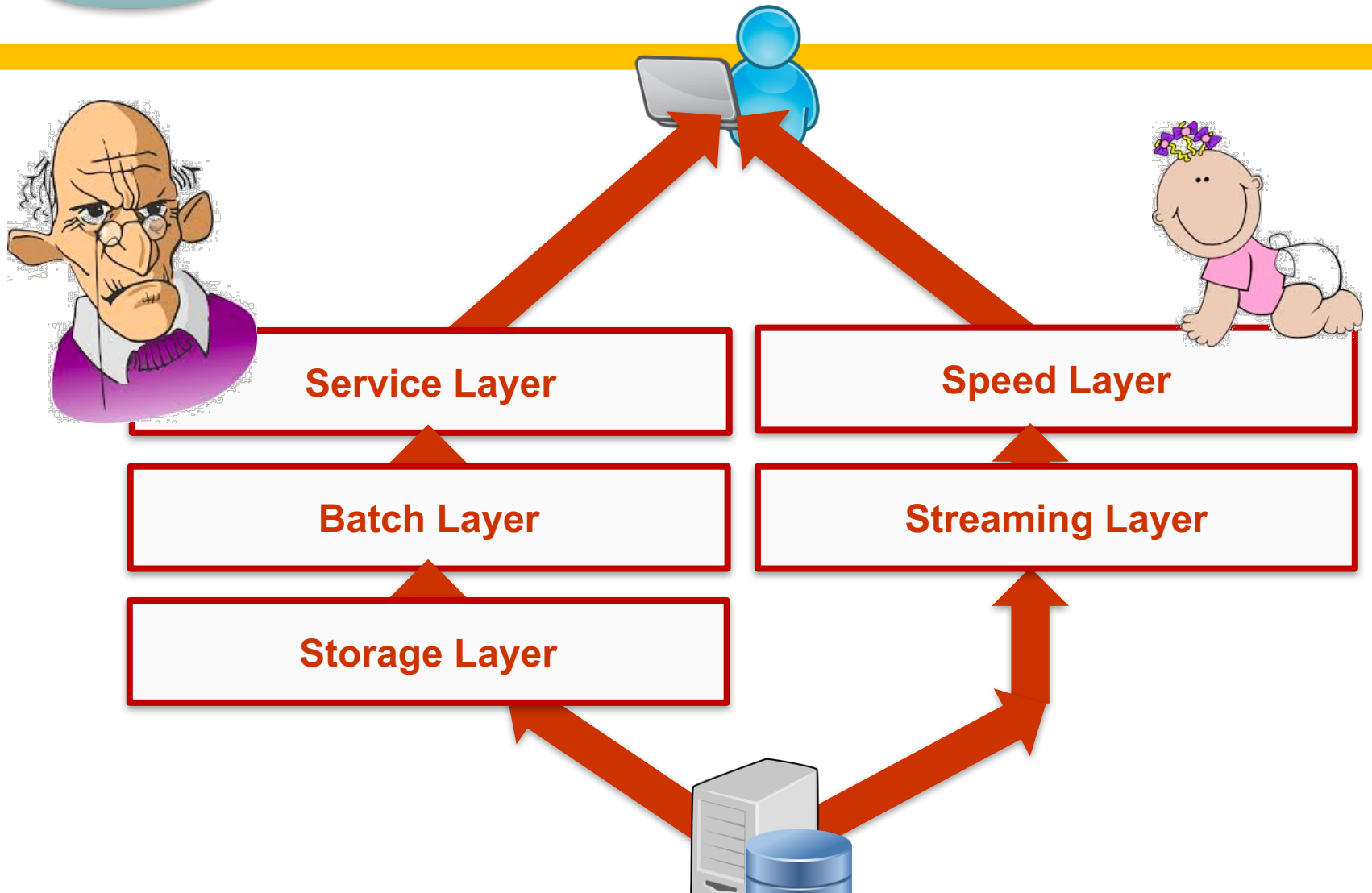
Big Data Framework: Lambda Architecture



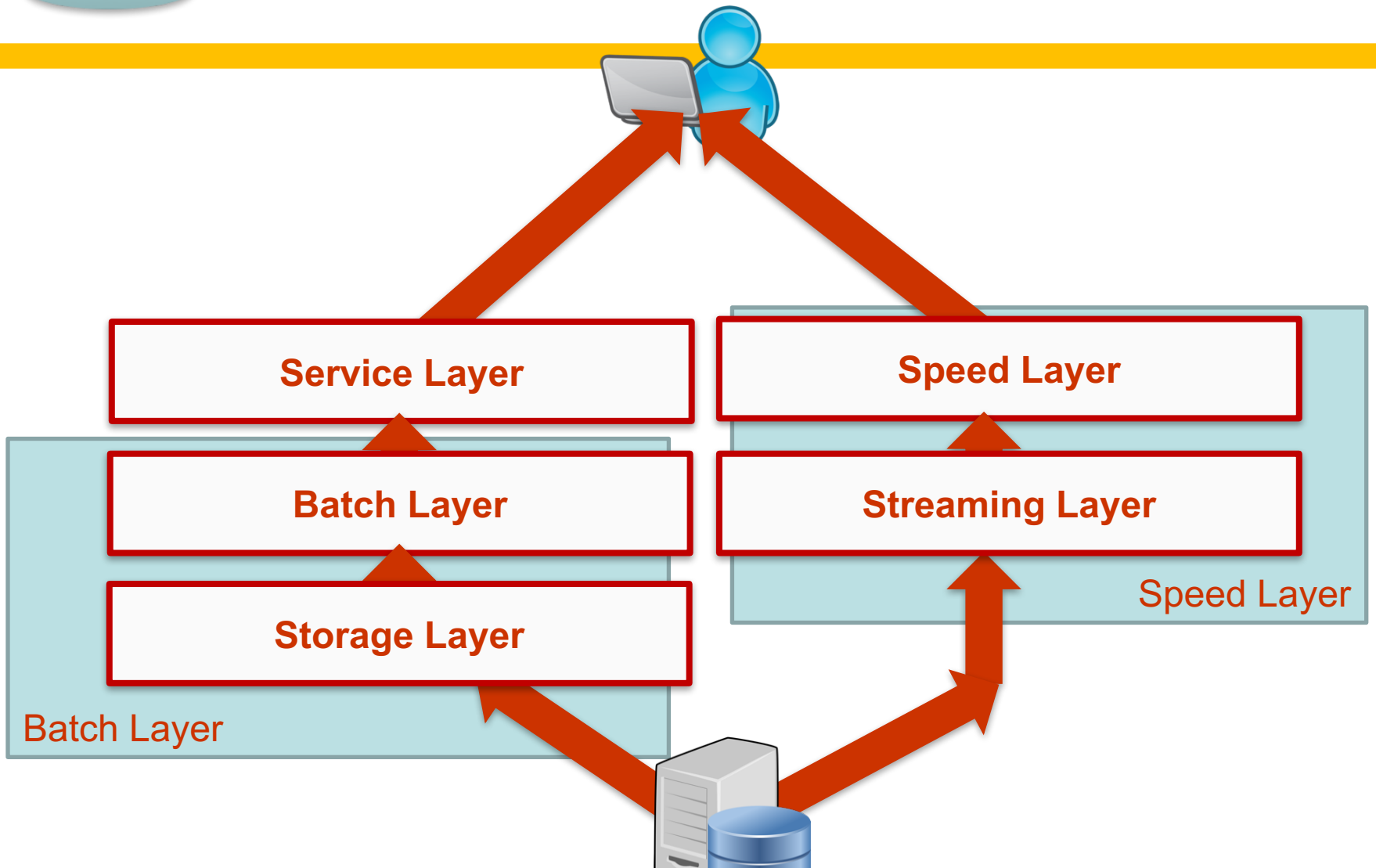
Data Streaming

- Data base =
  - Data at rest
  - Queries on the move
- Data stream =
  - Queries at rest
  - Data on the move

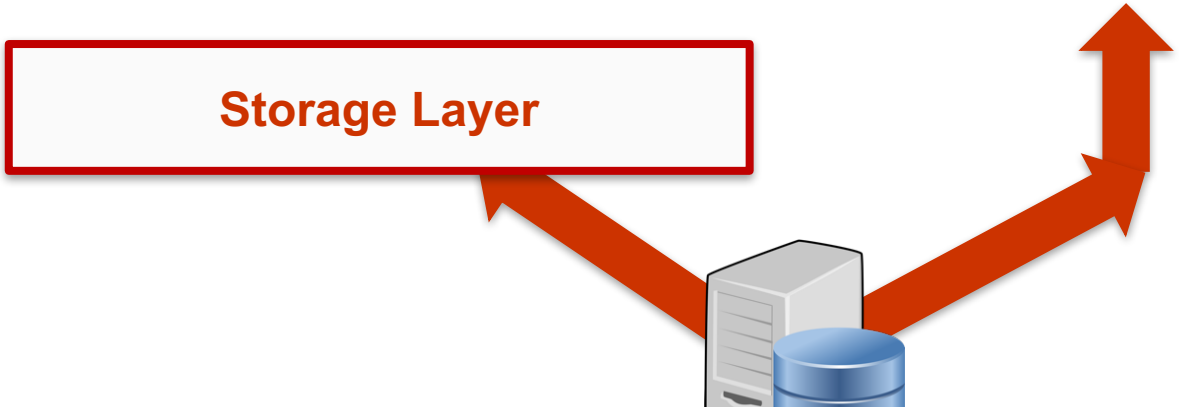
Big Data Framework: Lambda Architecture



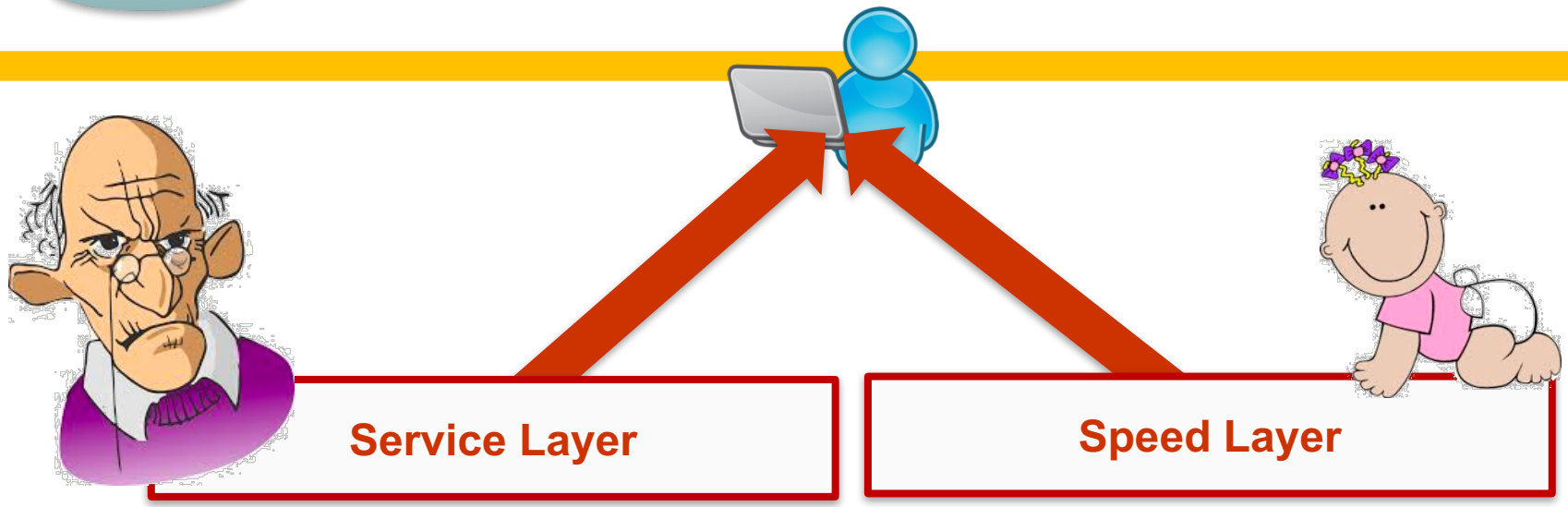
Big Data Framework: Lambda Architecture



Splitting Data



Joining Results

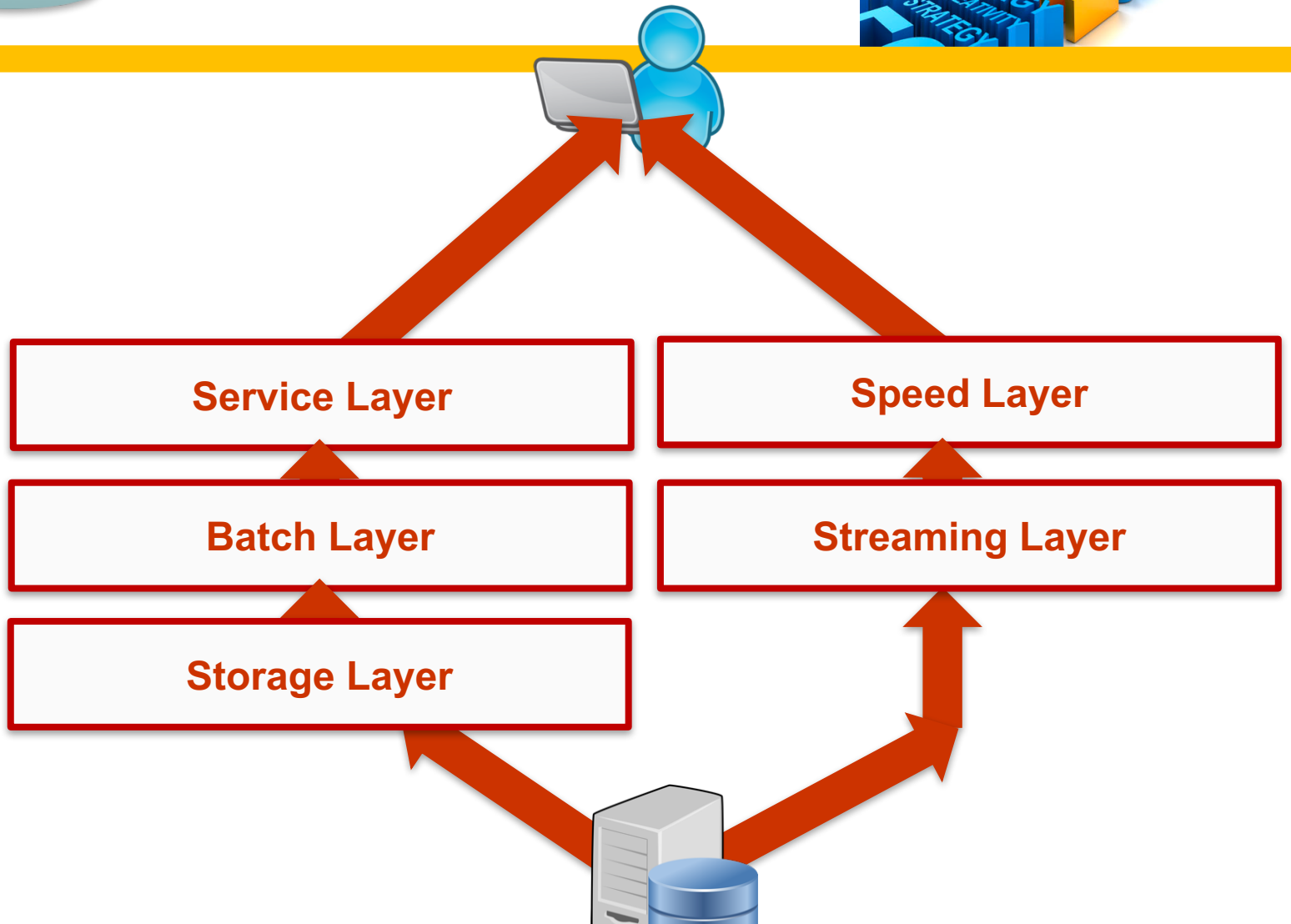


Lambda Architecture vs Requirements

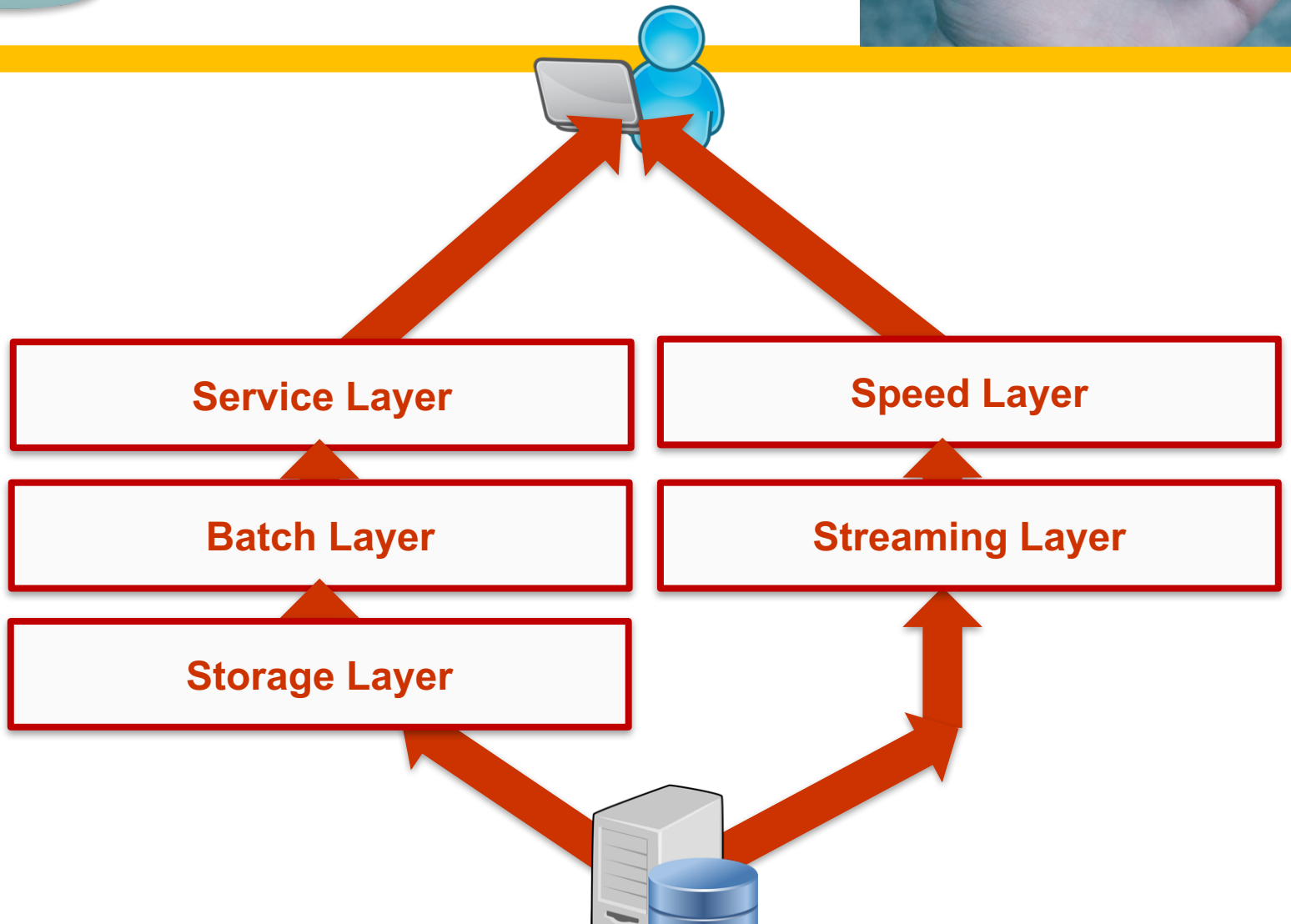
IT University of Copenhagen

- Scalability?
- Low latency?
- Fault tolerance?
- Maintainability?

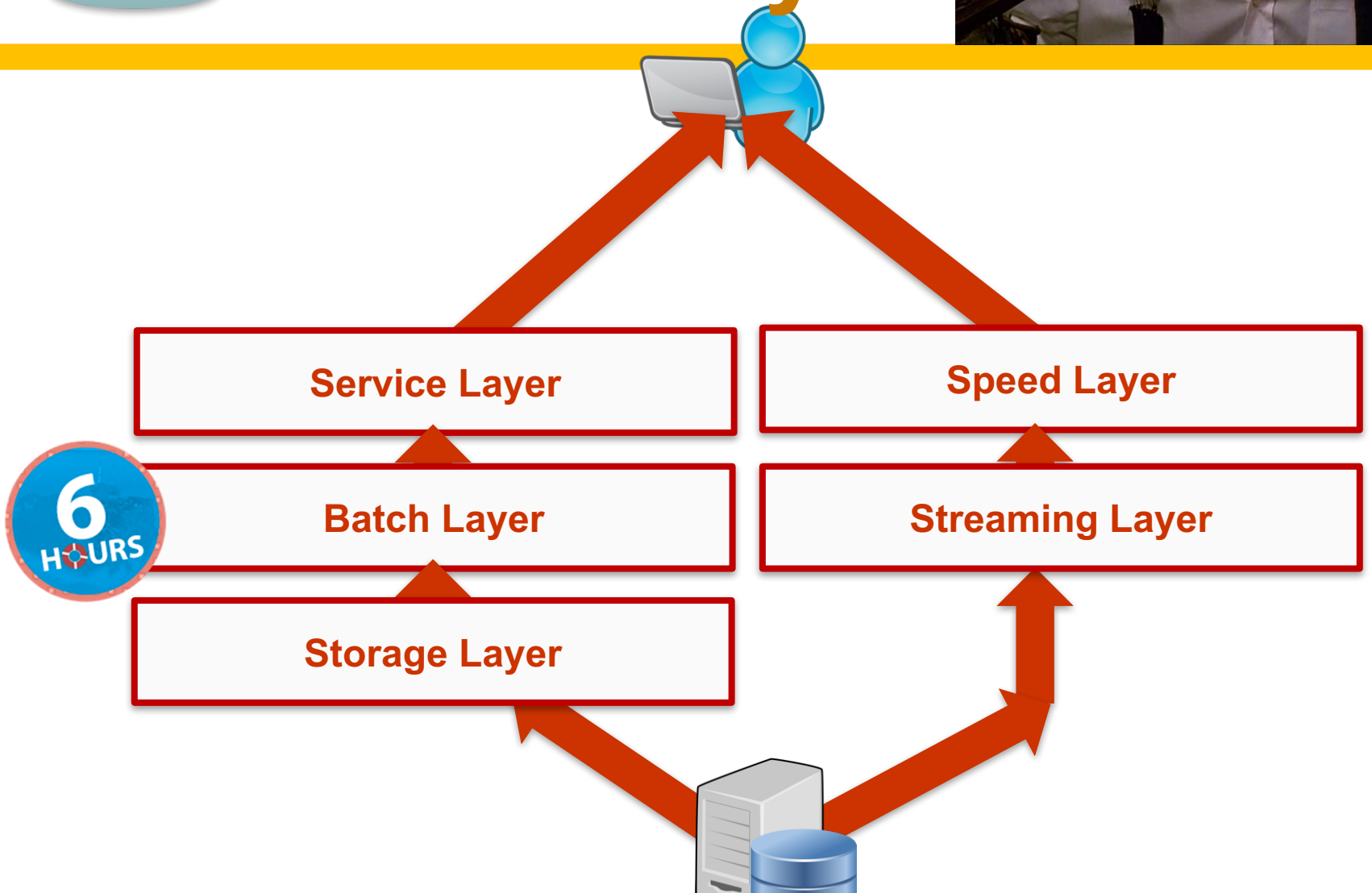
Scalability?



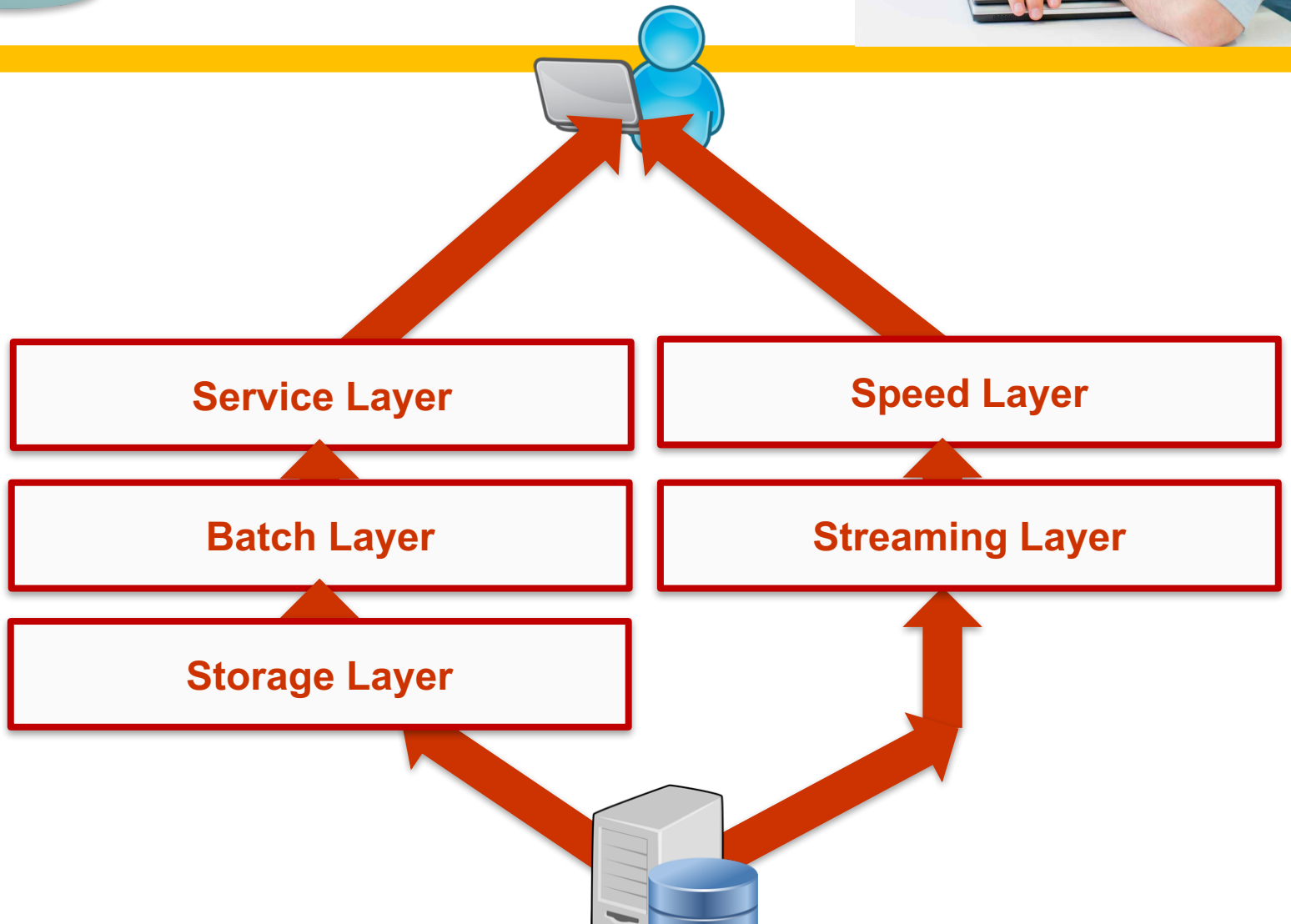
Latency? Read/Write



Failure Recovery



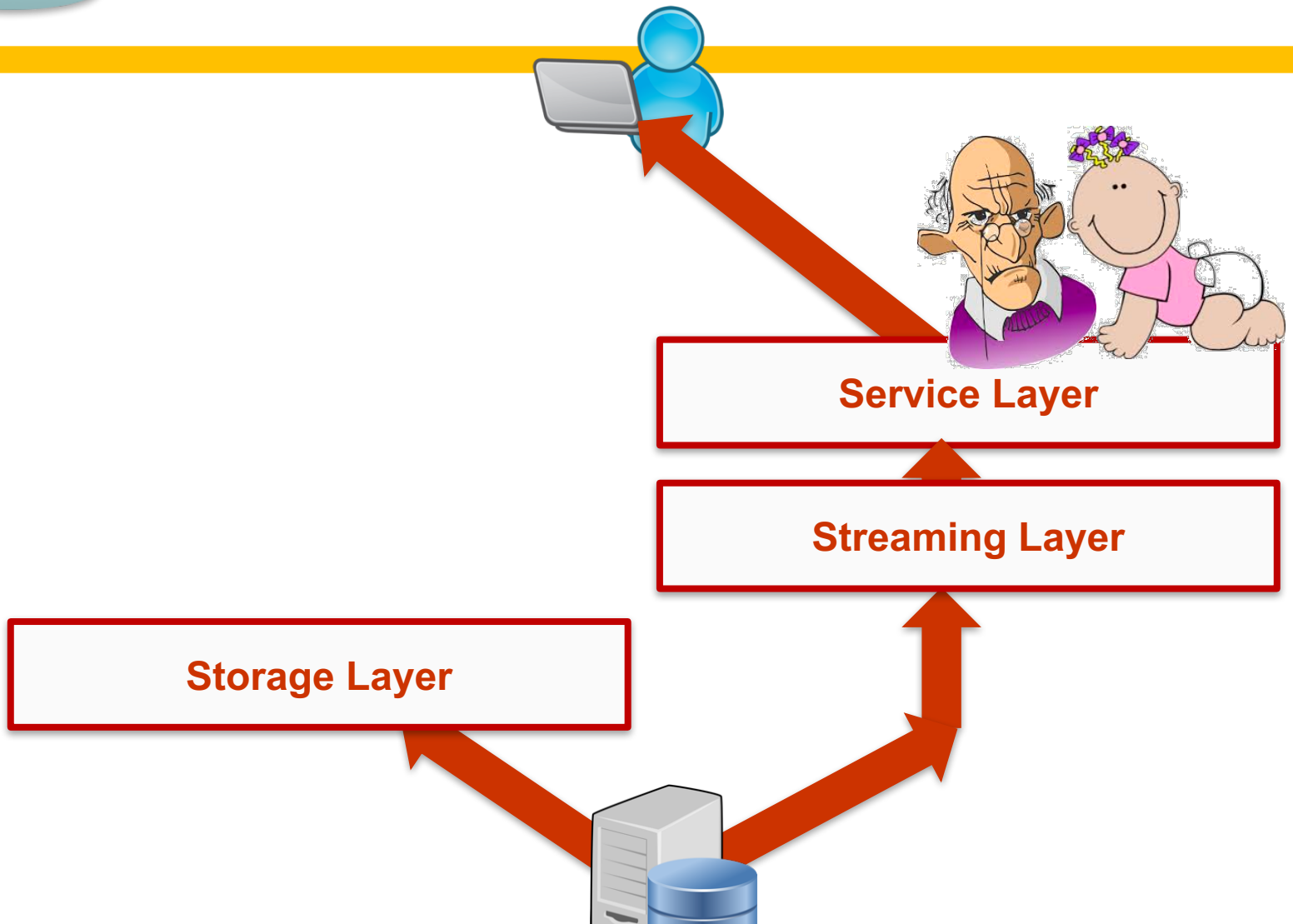
Easy to Maintain?



Criticism

- Not really an “architecture”
- Batch processes repeated ad infinitum
  - Wasted effort → incremental batches BUT…
- Cannot catch all errors!
- Joining results is cumbersome
- Two code bases
  - Two Service Layers
  - Batch Layer and Stream Layer

Kappa Architecture



Why Here?
Vehicle for Coverage…

Calendar Week	Calendar Date	Course Week	Reading Materials	Lecture Topics	Exercise Topics
35	29.08	1	Papers	TECH: Relational model; DB systems; Big data motivation	
36	05.09	2	Papers	CRIT: Raw vs cooked data; Data valence; Personal data; Data walk	CRIT: Data walk
37	12.09	3	Papers	CRIT: Social aspects; Immutability and certainty; Perils of interpretation and bias	CRIT: Bias Project 1
38	19.09	4	Ch 1-3	TECH: Lambda Architecture: Overview; Layers and roles; Data properties	TECH: Intro to Spark Shell
39	26.09	5	M&W: Ch 4-5	Batch Layer: Storage; High volume; DHT	Project 1
40	03.10	6	Papers	Intro do Data Cleaning CRIT: Andrew Clement guest lecture	Project 2
41	10.10	7	M&W: Ch 6-9	Batch Layer: Batch Processing; Spark	Spark; HDFS; Cluster
42					
43	07.01	8	M&W: Ch 10-13	Serving/Speed Layer: NoSQL and newSQL; CAP theorem	NoSQL
44	14.01	9	Papers: TBA	Data Pipeline Management; Metadata	Project 2
45	21.01	10	M&W: Ch 14-17	Speed Layer: Data streaming	Streaming Project 3
46	28.01	11	Papers: TBA	Data quality; Data integration	Data quality
47	04.02	12	Papers: TBA	TECH: OLAP & Data Mining 101	TECH: Intro to DM with Spark
48	11.02	13	M&W: Ch 18 + Papers: TBA	Lambda Architecture: Summary; Extensions	Project 3
49	18.02	14			Portfolio Work
50	25.02	15			Portfolio Work
				Exam paper due 18.12