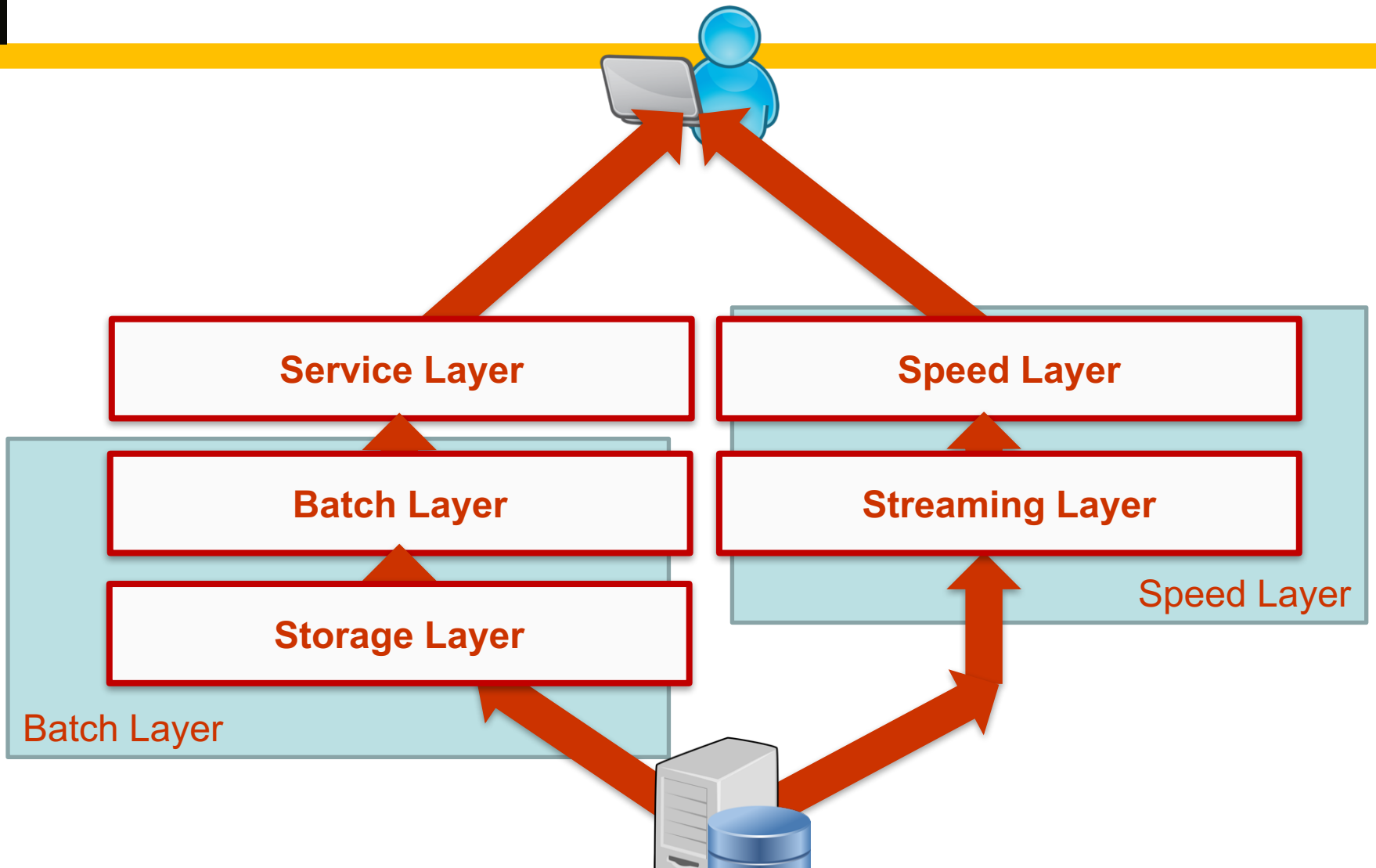
IT UNIVERSITY OF COPENHAGEN

Big Data Management
Pipeline Management

Björn Þór Jónsson

Lambda Architecture



Part I: Metadata

For something to exist, it has to be observed.

For something to exist, it has to have a position in time and space.

And this explains why nine-tenths of the mass of the universe is unaccounted for.

Nine-tenths of the universe is the knowledge of the position and direction of everything in the other tenth.

Every atom has its biography, every star its file, every chemical exchange its equivalent of the inspector with a clipboard.

It is unaccounted for because it is doing the accounting for the rest of it, and you cannot see the back of your own head.

Nine-tenths of the universe, in fact, is paperwork.

Terry Pratchett, Thief of Time

What is Metadata?

- Data about data – **paperwork**
  - "Metadata is defined as the data providing information about one or more aspects of the data; it is used to summarize basic information about data which can make tracking and working with specific data easier."
- Descriptive / Structural / Administrative

https://en.wikipedia.org/wiki/Metadata

Descriptive Metadata

- Describes a resource for purposes such as discovery and identification
  - Title
  - Abstract
  - Author
  - Keywords

https://en.wikipedia.org/wiki/Metadata

Structural Metadata

- Metadata about containers of data and indicates how compound objects are put together, for example, how pages are ordered to form chapters
  - Types
  - Versions
  - Relationships

https://en.wikipedia.org/wiki/Metadata

Administrative Metadata

- Provides information to help manage a resource
  - When and how it was created
  - File type and other technical information
  - Who can access it

https://en.wikipedia.org/wiki/Metadata

Examples I

- *Descriptive / Structural / Administrative*

- Book: “Big Data” by Marz & Warren
- Your latest programming assignment

- A sales database
- A sales record in a database

Examples II

- *Descriptive / Structural / Administrative*

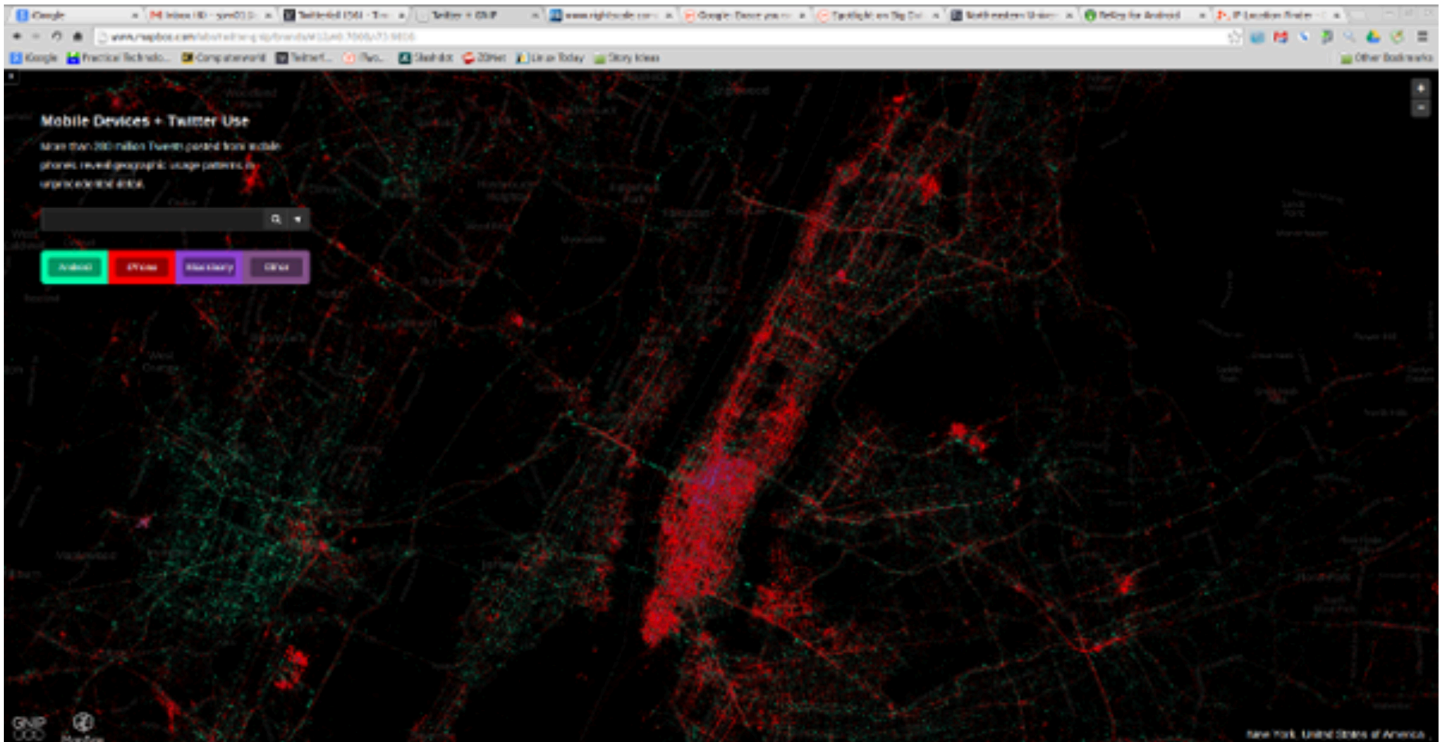
- The Europeana Collection
- Night Watch by Rembrandt
- Photo of Night Watch stored by Europeana

Metadata Storage

- Internal (embedded)
- External (database/repository)

Metadata vs Data

- *Big data, metadata, and traffic analysis: What the NSA is really doing*



Importance of Metadata

- *Why you need metadata for Big Data success*
  - … provides the value and purpose of the data content
  - … effective tool for quickly locating information
  - … can link your firm's data assets
  - … create and maintain data consistency
- Metadata and the 'Big Data Gap'

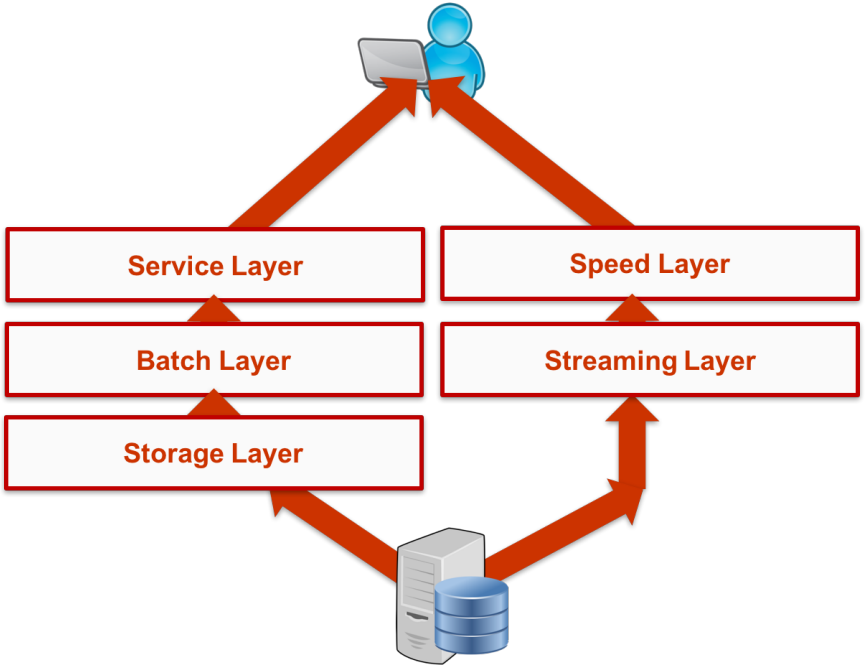
Importance of Metadata

SMALL DATA		BIG DATA
Specific questions	***GOAL***	Broad concerns
One location	***LOCATION***	Many locations
Structured	***STRUCTURE***	Varied, unstructured
Single user	***SOURCE***	Many providers
Transient	***LONGEVITY***	Durable
Focused	***MEASUREMENTS***	Broad
Can be recreated	***REPRODUCIBILITY***	Gone if not captured
Small risk	***STAKES***	Big risk
Simple	***INTROSPECTION***	Metadata is vital
Complete	***ANALYSIS***	Incremental

Source: Jules J. Berman, Big Data Principles

Data Lineage and Traceability

- A client sees a strange record – how did it come into existence?



GDPR

- *EU regulations on algorithmic decision-making and a "right to explanation"*
- Right to knowledge, to be forgotten, ...
- Right to explanation
  - … a data subject has the right to "an explanation of the decision reached after [algorithmic] assessment."
  - … what does it mean, and what is required, to explain an algorithm's decision?

From Statistics to AI

- Linear regression, correlation
- Support vector machines
- Ensemble methods
- Neural networks
  - "what hope is there of explaining the weights learned in a multilayer neural net with a complex architecture?"

Technical vs Ethical

Above all else, the GDPR is a vital acknowledgement that, when algorithms are deployed in society, few decisions if any are purely "technical". Rather, the ethical design of algorithms requires coordination between technical and philosophical resources of the highest caliber. A start has been made, but there is far to go. And, with less than two years until the GDPR takes effect, the clock is ticking.

Metadata in Lambda Architecture

Metadata in Lambda Architecture

- Storage layer contains some metadata
  - Is it enough?
  - What about the Service Layer?

- Advocates higher-level abstractions
  - Implicit metadata
  - Is it enough?

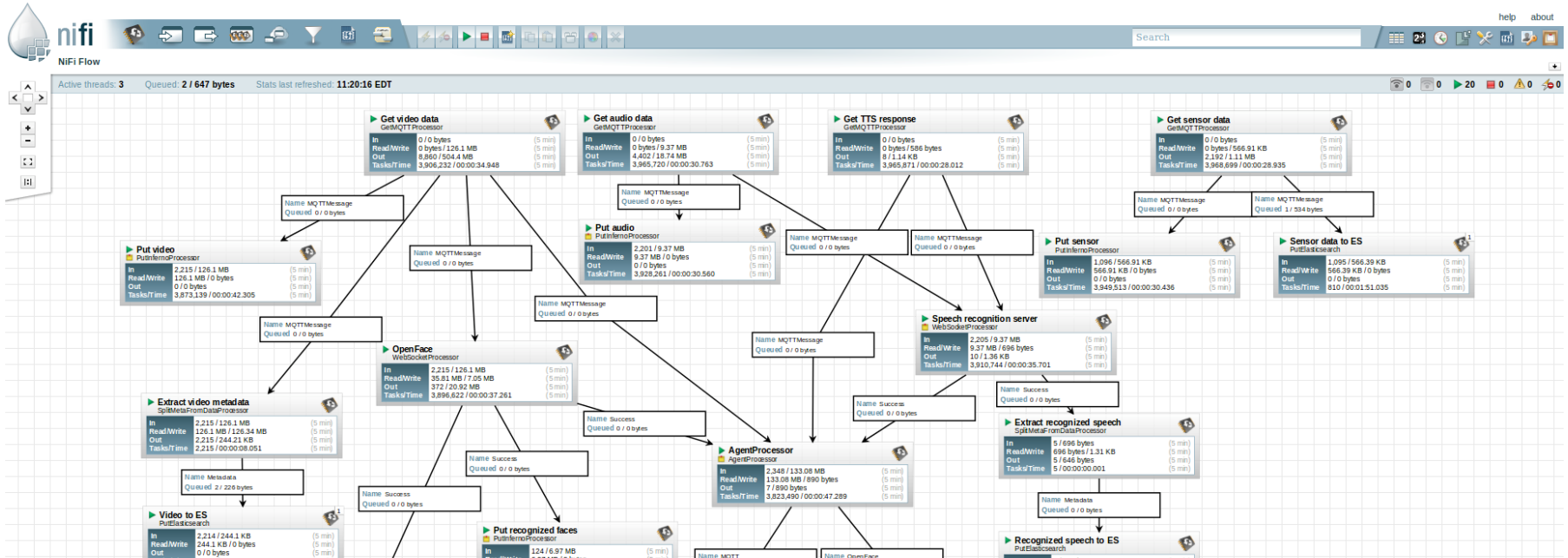
Part II: Pipeline Mgmt

Pipeline Orchestration

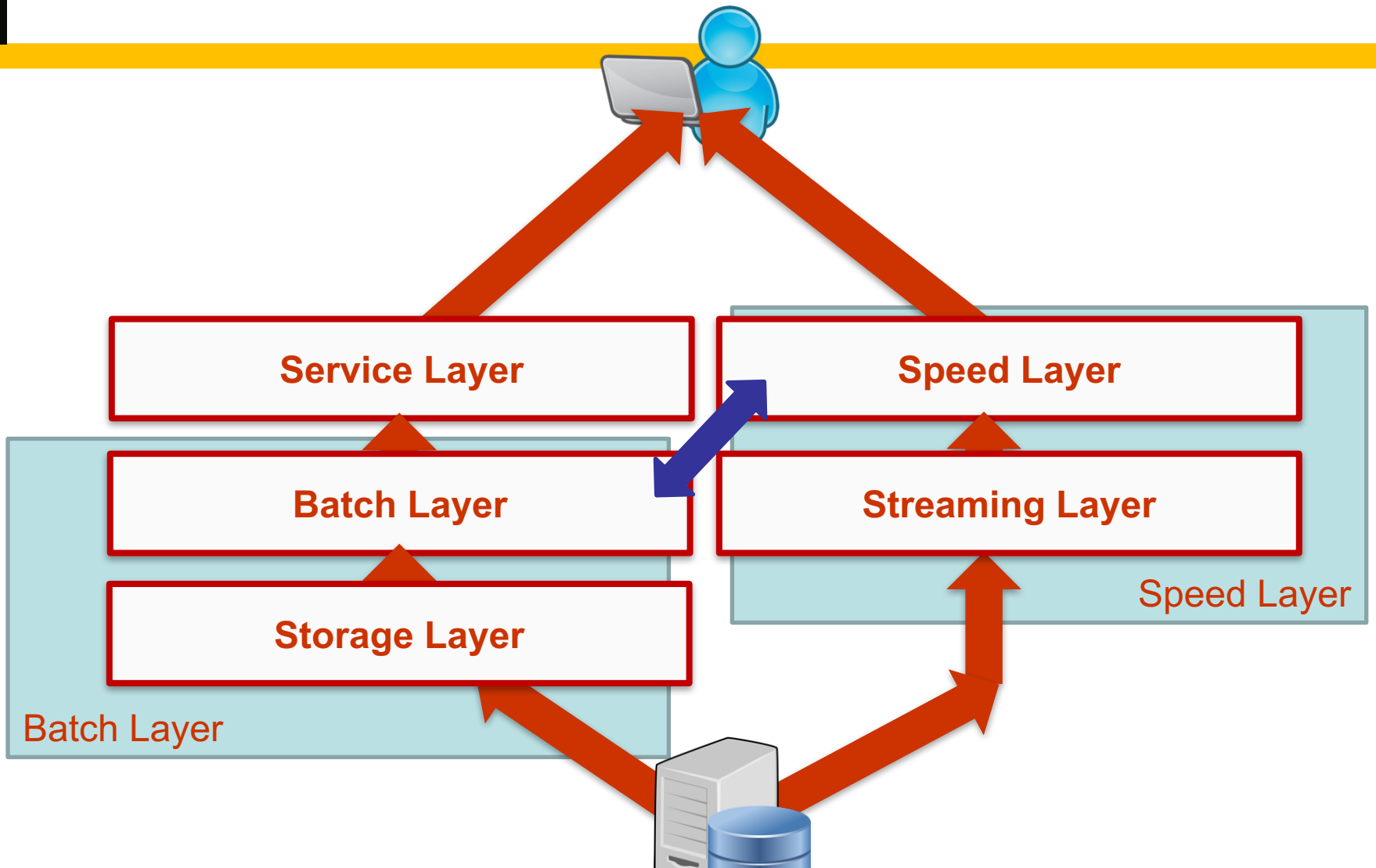
- So far
  - Execution packages
  - Scheduled jobs (cron)
- Tools?
  - Graphical tools – ETL – data flow
  - High-level abstraction and synchronization
- Key: Traceability!

Apache Nifi

- Flow Based Programming
  - Each flow is a stream with associated metadata (KV pairs)
  - DAG of processors
- Flow Management
  - Guaranteed Delivery
  - Buffering
  - Prioritized queuing
- Provenance is maintained



Lambda Architecture



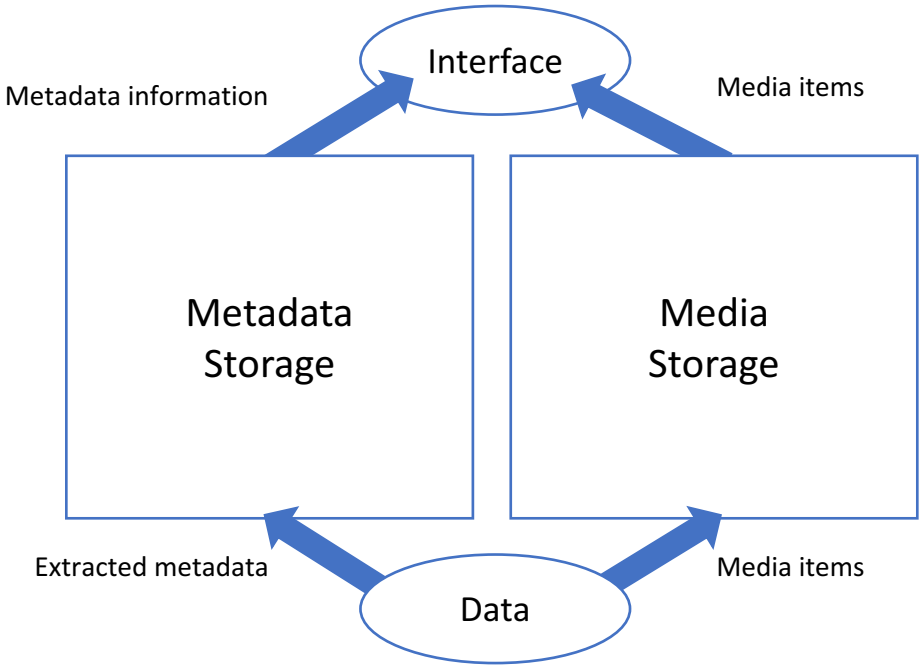
Re-Computation vs Incremental Views

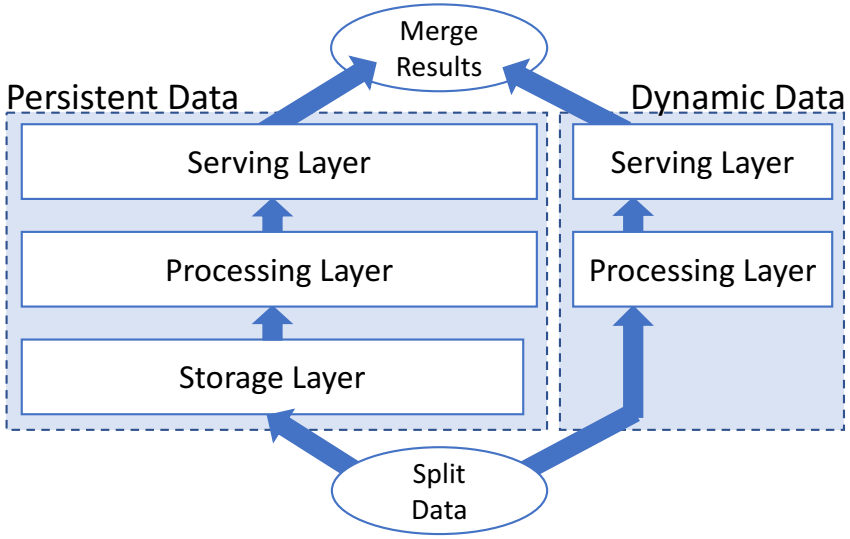
- Re-Computation
  - Simple
  - Must be an option
- Incremental Views
  - Faster, more resource-friendly
  - More complex, may need additional data
  - May have levels of incremental views

Layered Layers

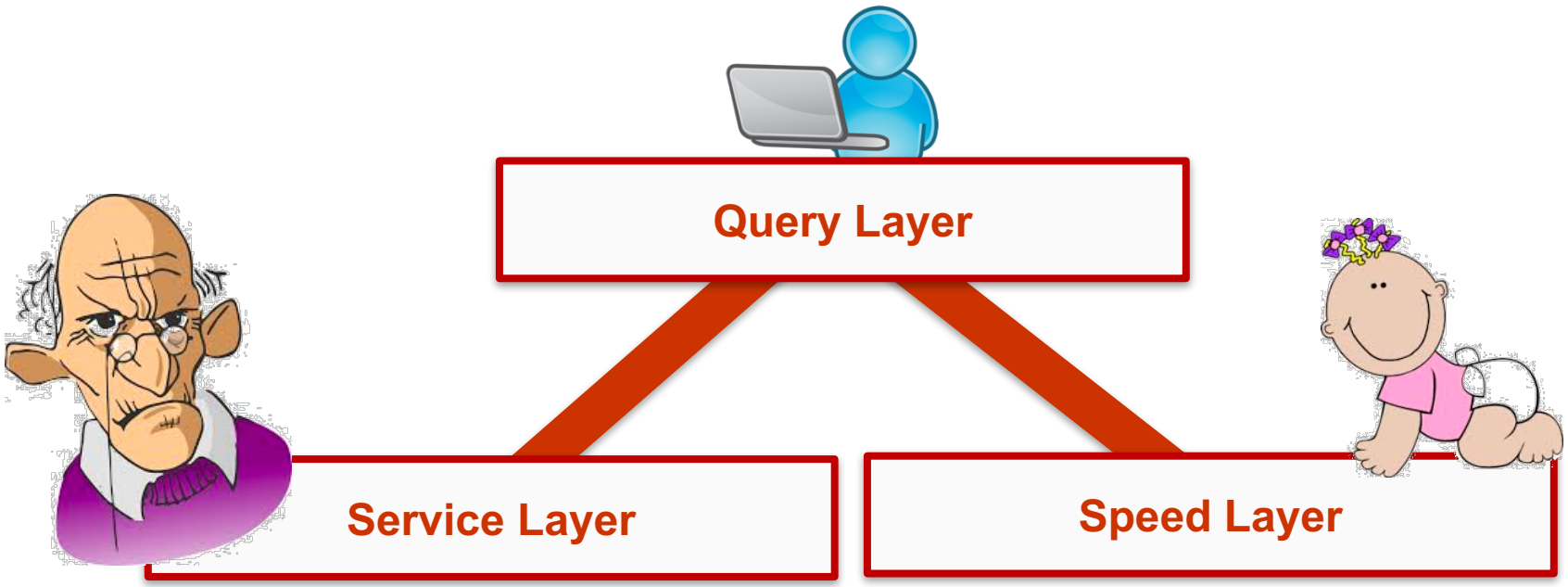
- Some batch computations may be shared
  - Intermediate views – avoid computing > 1x
- Propose layered Batch Layer

Big Multimedia





Query Layer



Take Away

- Need to seriously address metadata
  - GDPR requirements, lineage, traceability
  - Descriptive / Structural / Administrative
  - How to store/manage metadata?
- Pipeline management
  - Need to orchestrate the workflow
  - Lambda Architecture can be tweaked!