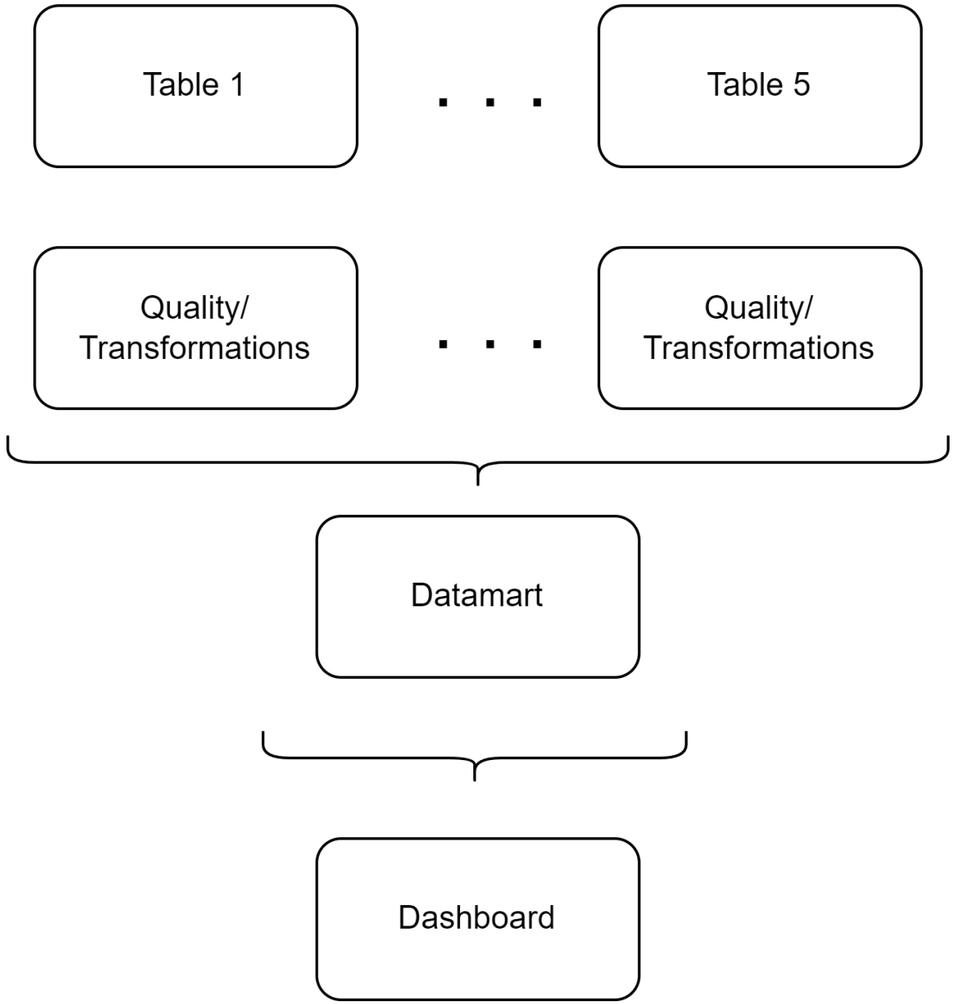
Orchestration

UNDERSTANDING MODERN DATA ARCHITECTURE

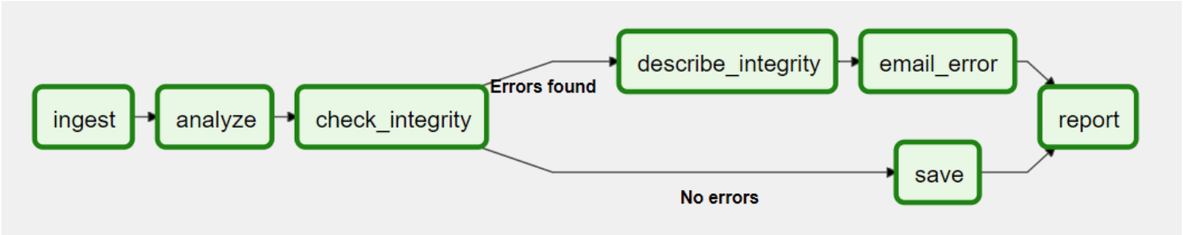
Miller Trujillo
Senior Software Engineer

datacamp

What is orchestration?



- Coordinate multiple jobs
- Automated configuration and coordination of complex workflows.



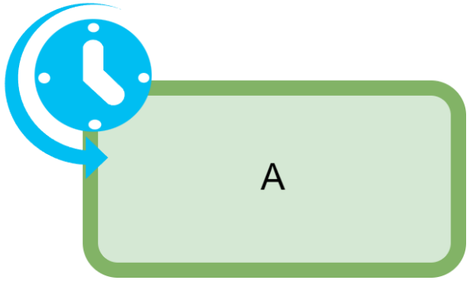
- **Frees up human resources**

[1] https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/dags.html

Orchestration vs scheduling

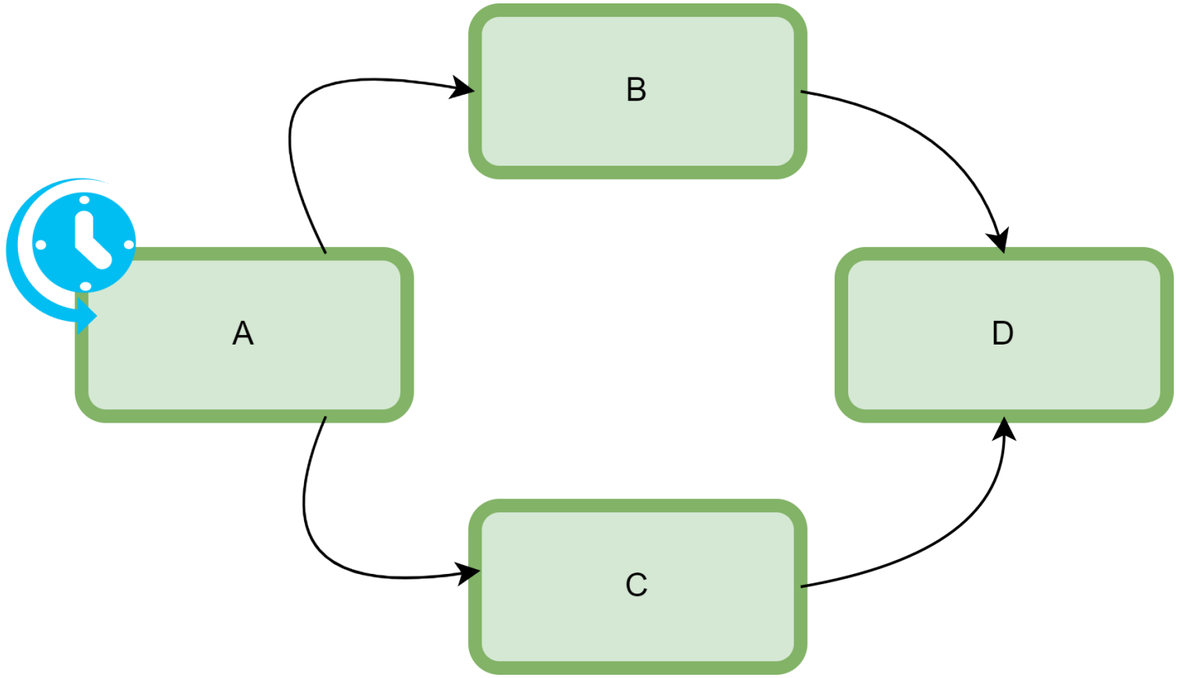
Scheduling

- Execute tasks at specified intervals/times
- Starter of orchestrated workflows



Orchestration

Automate and coordinate complex workflows



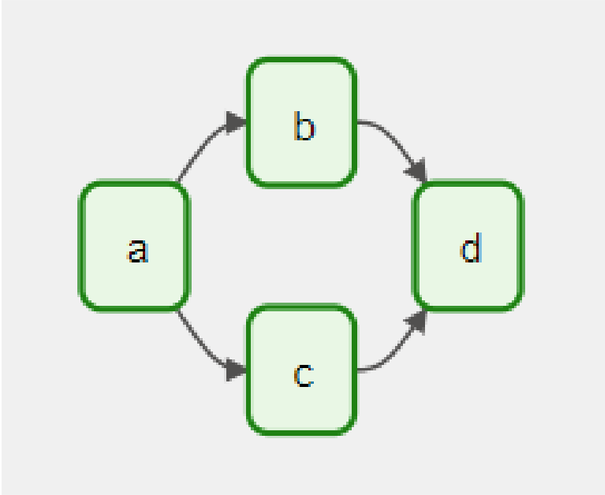
Apache Airflow

Apache Airflow

- Coordinate complex workflows with **Python**

Core concepts of orchestration

- Tasks:
  - Basic unit of execution
- Dependencies:
  - Determine task sequence
- Directed Acyclic Graph (DAG):
  - Workflow of tasks and dependencies



Core concepts of orchestration

- Operators:
  - Determine nature of task
  - `BashOperator`
  - `PythonOperator`
- Sensors:
  - Wait for specific conditions
- Scheduler
  - Automates triggering of tasks.

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Storage & processing costs best practices

UNDERSTANDING MODERN DATA ARCHITECTURE

Miller Trujillo
Senior Software Engineer

datacamp

Cost models

Pay-as-you-go

- Pay for what you consume
- Bytes stored
- Compute used per minutes used

Reserved capacity

- Plan and pay upfront
- Discounts!

Costs to keep in mind

Common charges

- Network traffic
- Bytes stored
- Compute capacity
- **Time of usage!!**

Blob storage cost example

Cloud storage

- Bytes stored
- Time bytes are stored
- Operations over data
  - Moving it

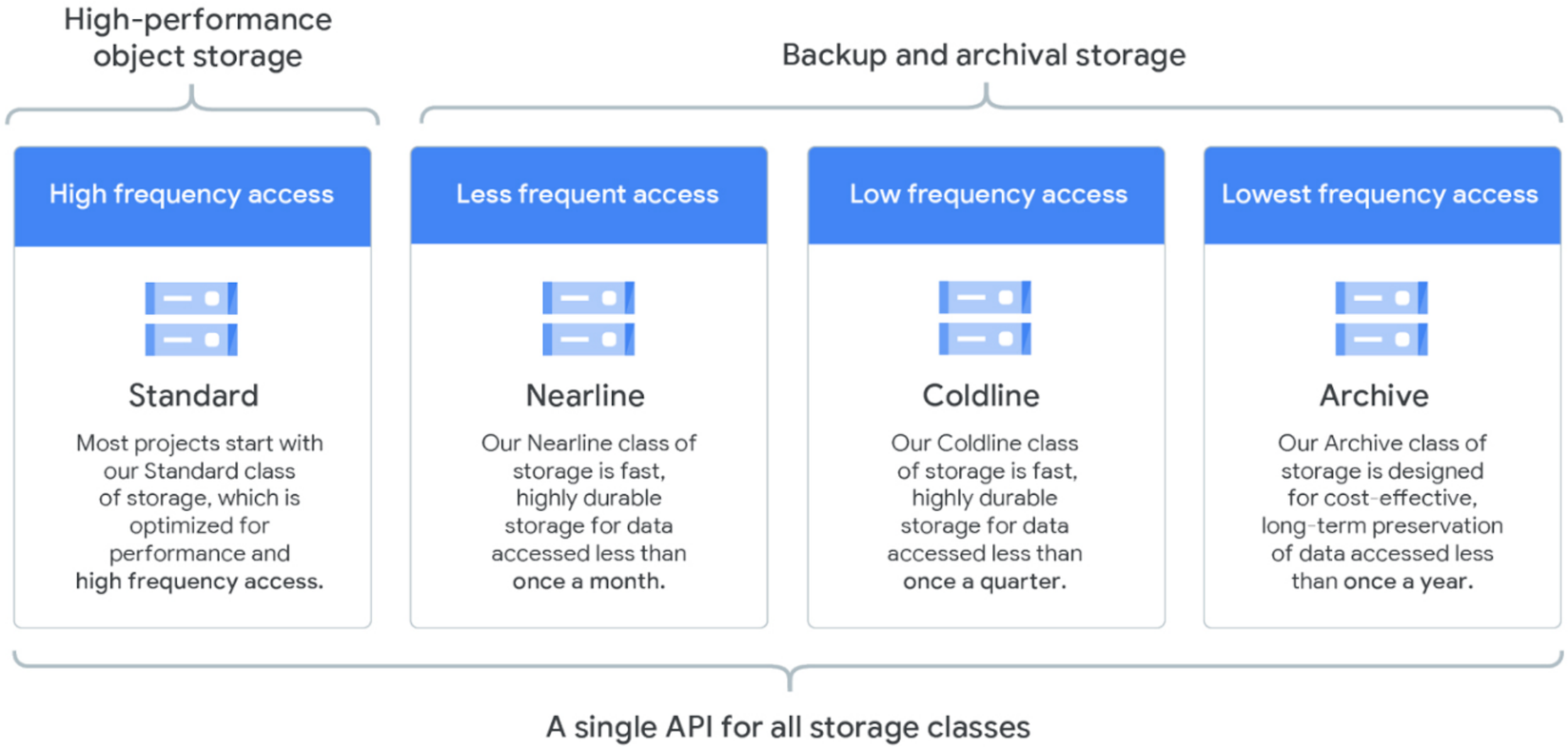
Store our data in regions close to users!!!

Networking

- Network traffic almost always is part of the charges!

Cost optimization

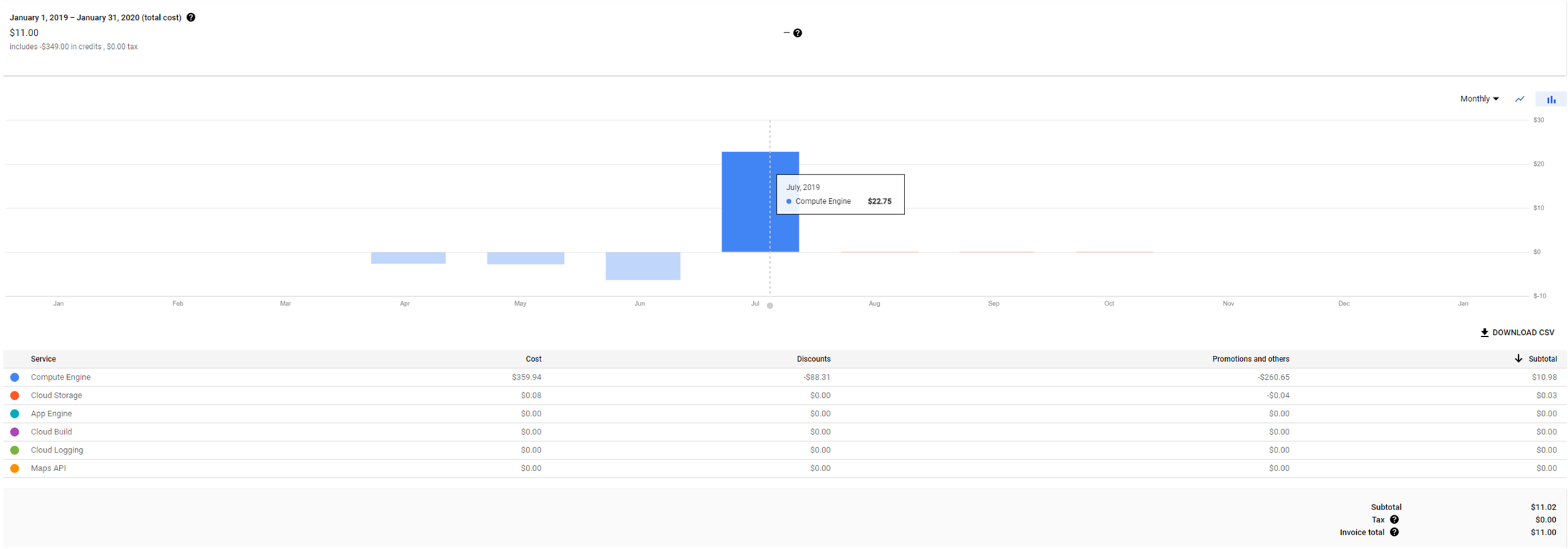
- Reduce costs while keeping the quality of service



[1] https://cloud.google.com/blog/products/storage-data-transfer/archive-storage-class-for-coldest-data-now-available

Importance of cost monitoring and alerting

- Continuously track cloud usage
- Notify unusual spending spikes
- Regularly review usage and adapt



Service	Cost	Discounts	Promotions and others	Subtotal
Compute Engine	$359.94	-$88.31	-$260.65	$10.98
Cloud Storage	$0.08	$0.00	-$0.04	$0.03
App Engine	$0.00	$0.00	$0.00	$0.00
Cloud Build	$0.00	$0.00	$0.00	$0.00
Cloud Logging	$0.00	$0.00	$0.00	$0.00
Maps API	$0.00	$0.00	$0.00	$0.00

Subtotal	$11.02
Tax	$0.00
Invoice total	$11.00

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Designing a modern data architecture

UNDERSTANDING MODERN DATA ARCHITECTURE

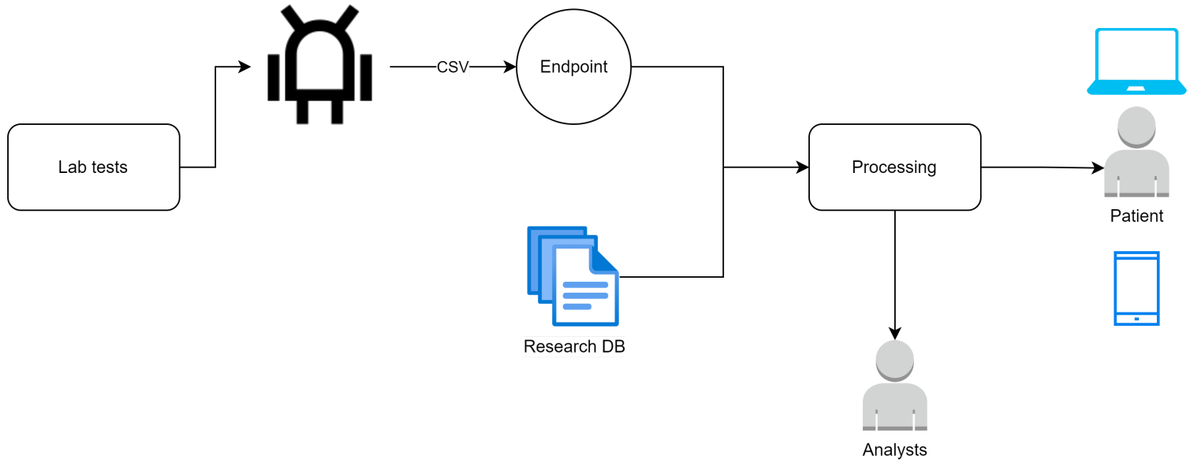
Miller Trujillo
Senior Software Engineer

datacamp

The business case

Medical laboratory

- Organizer robot
- Generate CSV
  - Up to 4 CSV every hour
- Databases in plain files



- Platform for patients to track results
- Enrich patients results with their investigation

Where to start?

Questions!

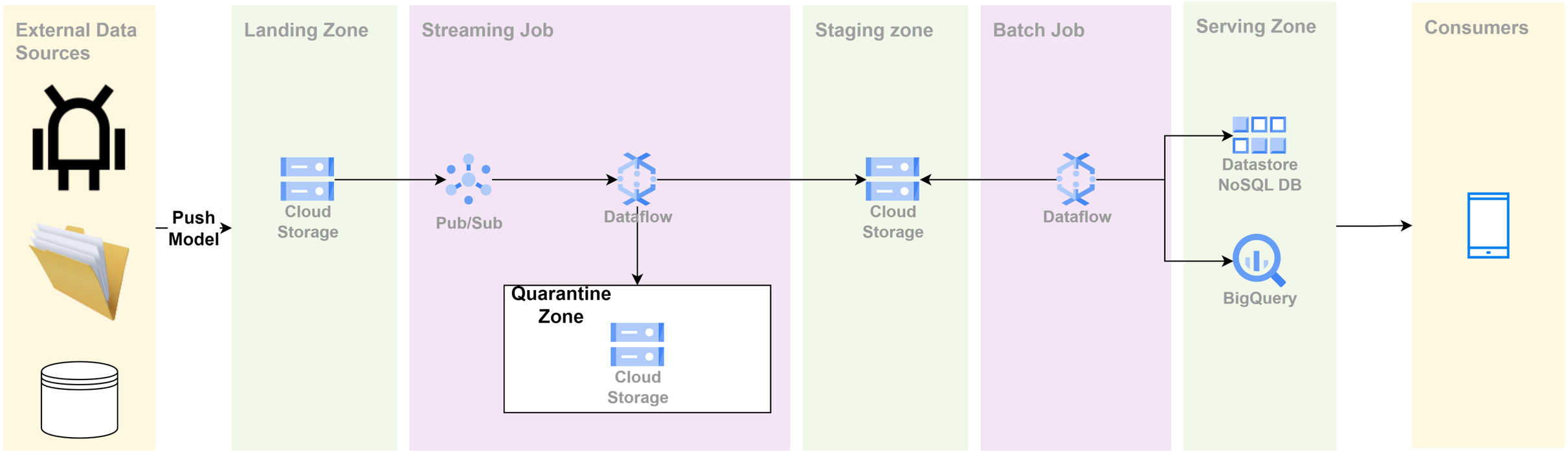
Refine the requirements!

- How large are the files?
- How many robots?
- How frequently are those files generated?
- How many files do they have?
- How data will be processed? Machine learning? Queries?
- How data will be exposed?
- What regulations/constraints do we have?

The assumptions

- 100 machines
- Each CSV file is around 100MB
- Plain files as database
- Tens of gigabytes for each plain file
- Model exposed through API
  - Requires all previous result or summary
  - Summary needs to be updated constantly
- Mobile app
- Ignore regulations

The solution



Let's practice!

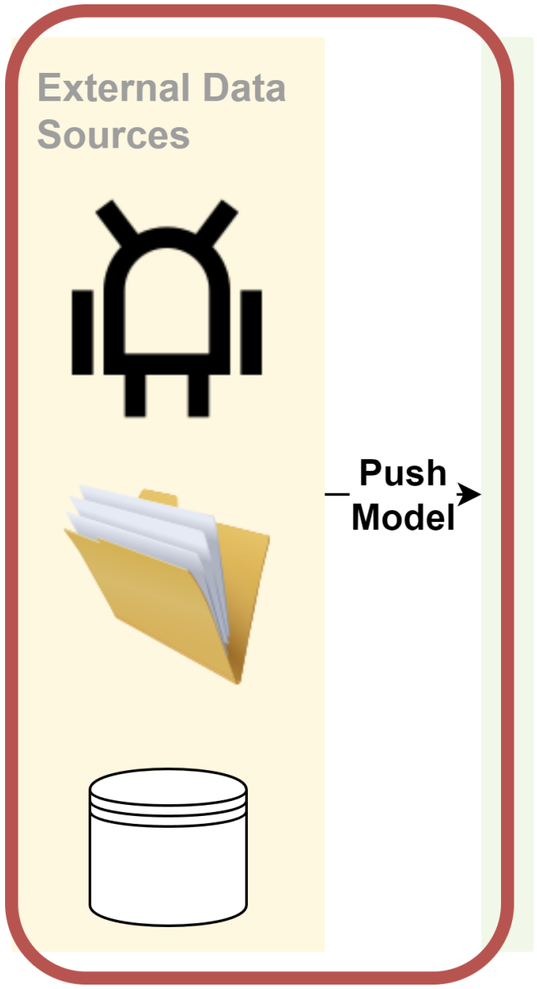
UNDERSTANDING MODERN DATA ARCHITECTURE

Evaluating modern data architecture solutions

UNDERSTANDING MODERN DATA ARCHITECTURE

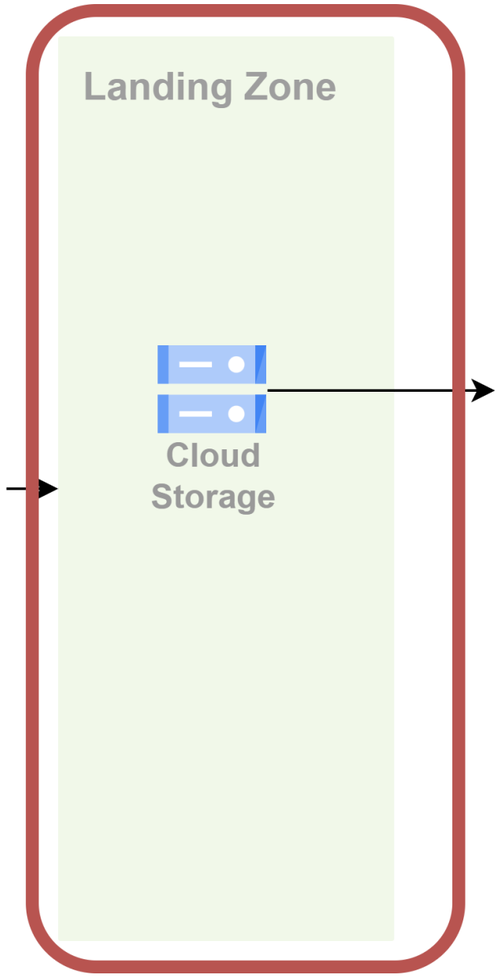
Miller Trujillo
Senior Software Engineer

Ingestion



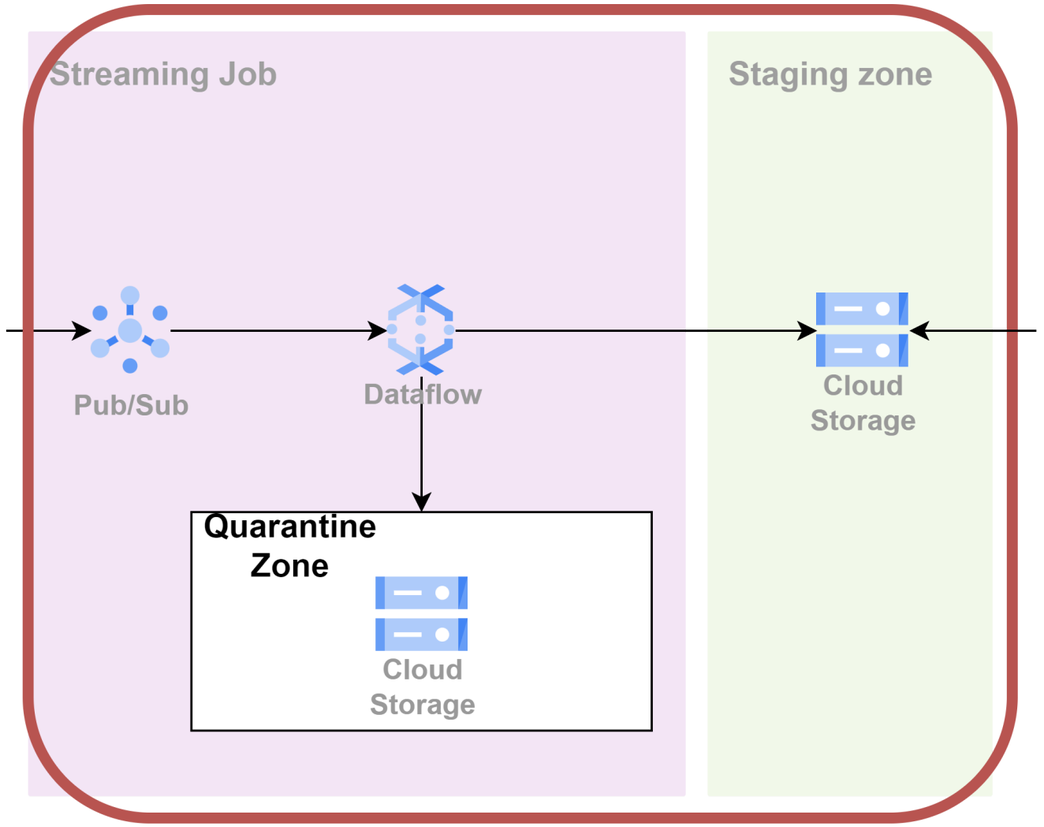
- Unpredictable patterns
- What if we pull the data?
  - Expose the files
  - Network file system

Storage



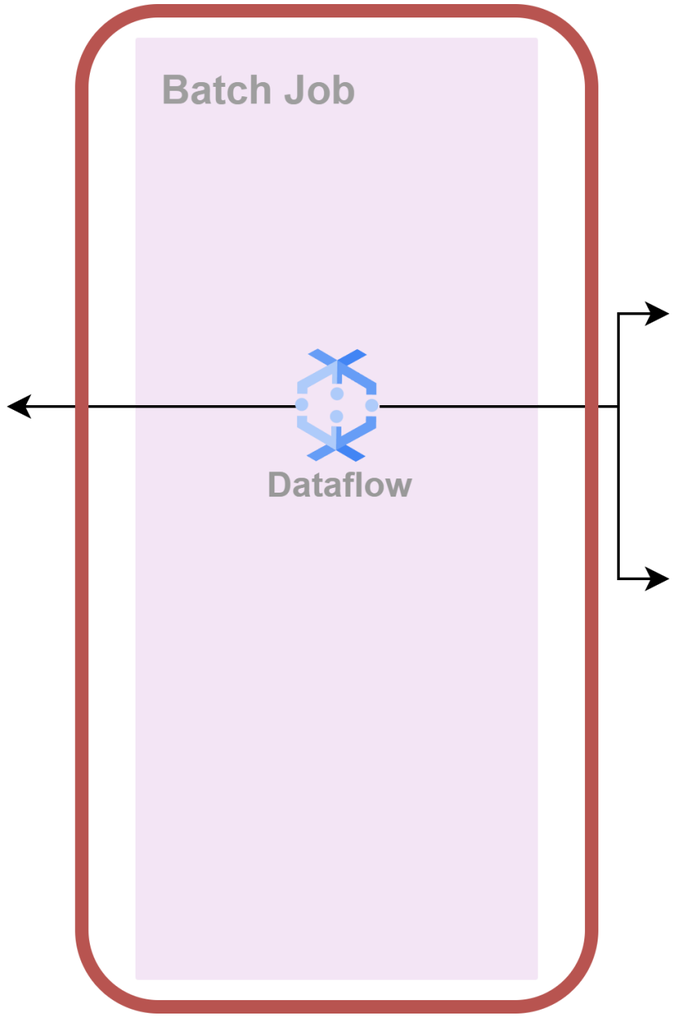
- Cloud storage is:
  - cheaper that data warehouse or databases
  - Flexible, and expose then required APIs
- BigQuery still an option?
  - Cheap enough
  - Not feasible due limitations at loading
- **Life-cycle policies to reduce even more the costs**

Processing



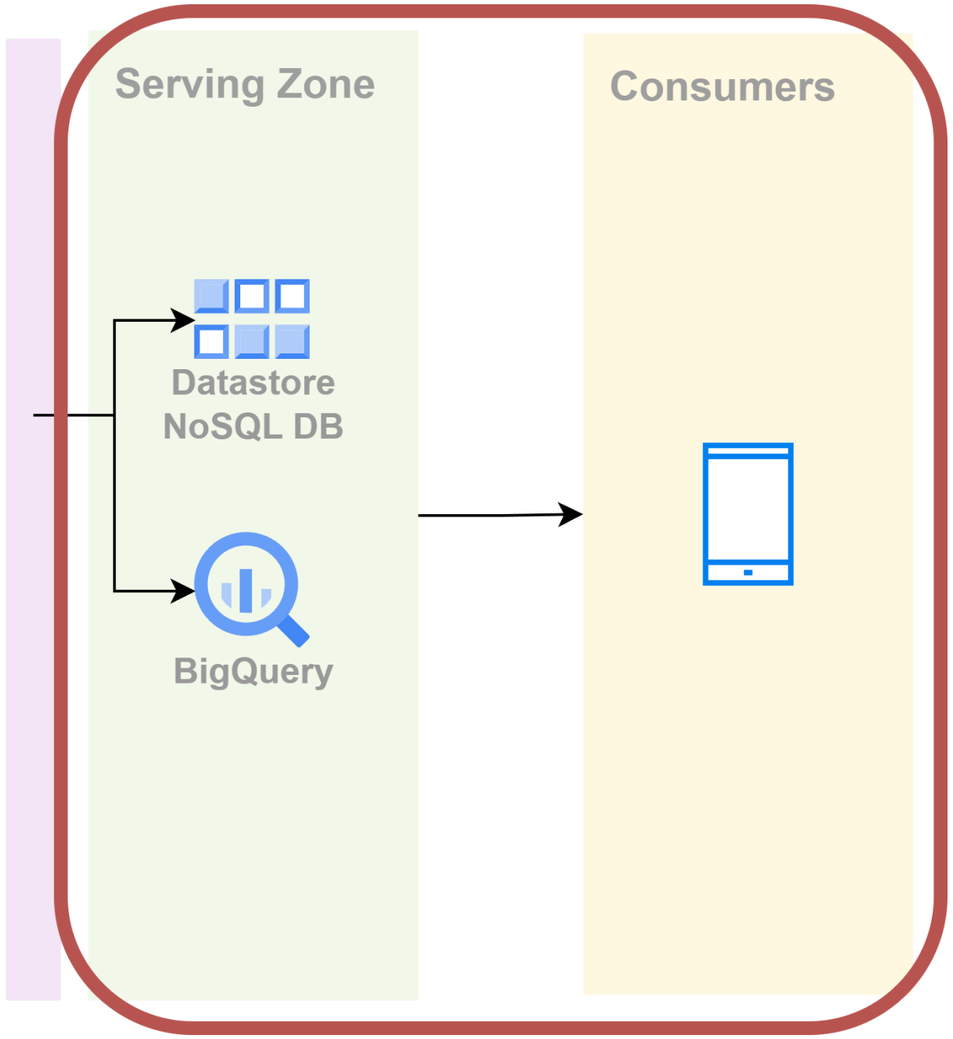
- Dataflow, Dataproc (Spark), or even Data Fusion
- Unpredictable arriving patterns
- Process data as soon as it arrives
- Simplicity
- Temporal data
- Automate cleaning with life-cycle policies
- No schema maintenance needed

Processing: The model scores



- Complex to keep track of everything
- Easier to maintain
- Previous job can write to NoSQL DB and this job complement data

Serving the data



- BigQuery for analytical purposes
- NoSQL DB => Easier scalability & flexibility

Some other details

- Governance, orchestration, security, among others
- Further refine the platform and requirements
- Enable better management
- Not one size fits all!

Everything is about trade-offs

Let's practice!

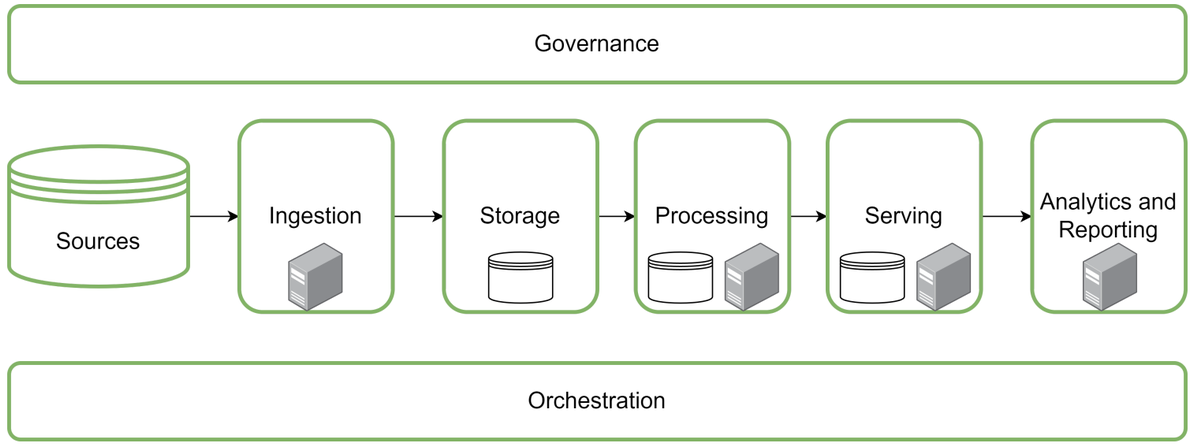
UNDERSTANDING MODERN DATA ARCHITECTURE

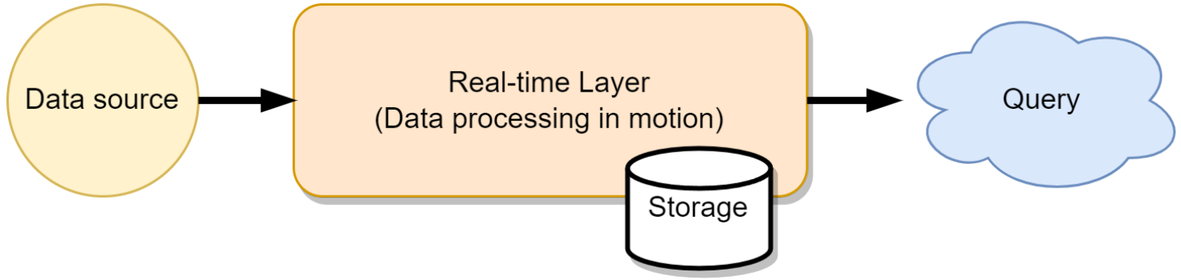
Wrap-up

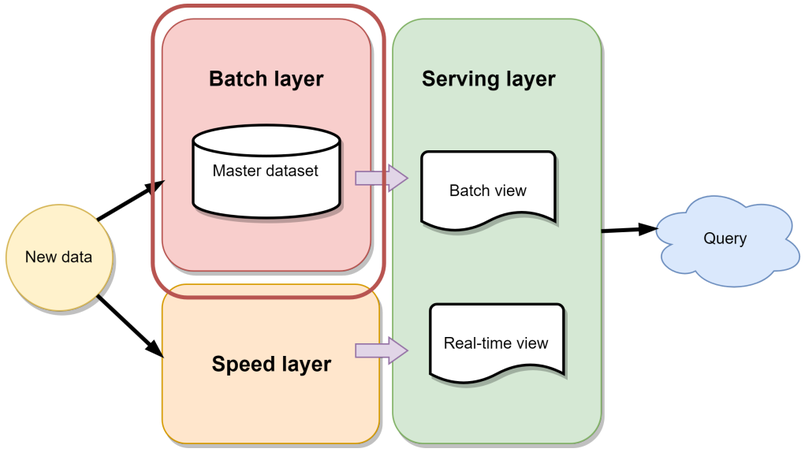
UNDERSTANDING MODERN DATA ARCHITECTURE

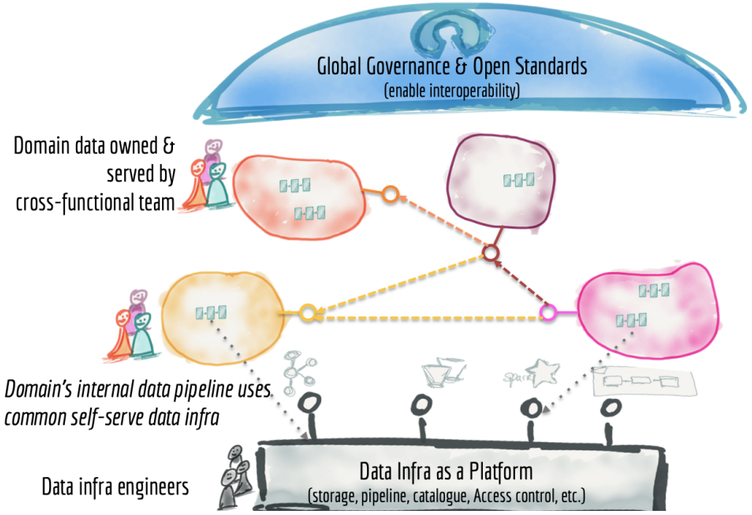
Miller Trujillo
Senior Software Engineer

Chapter 1 - Introduction to Modern Data Architecture



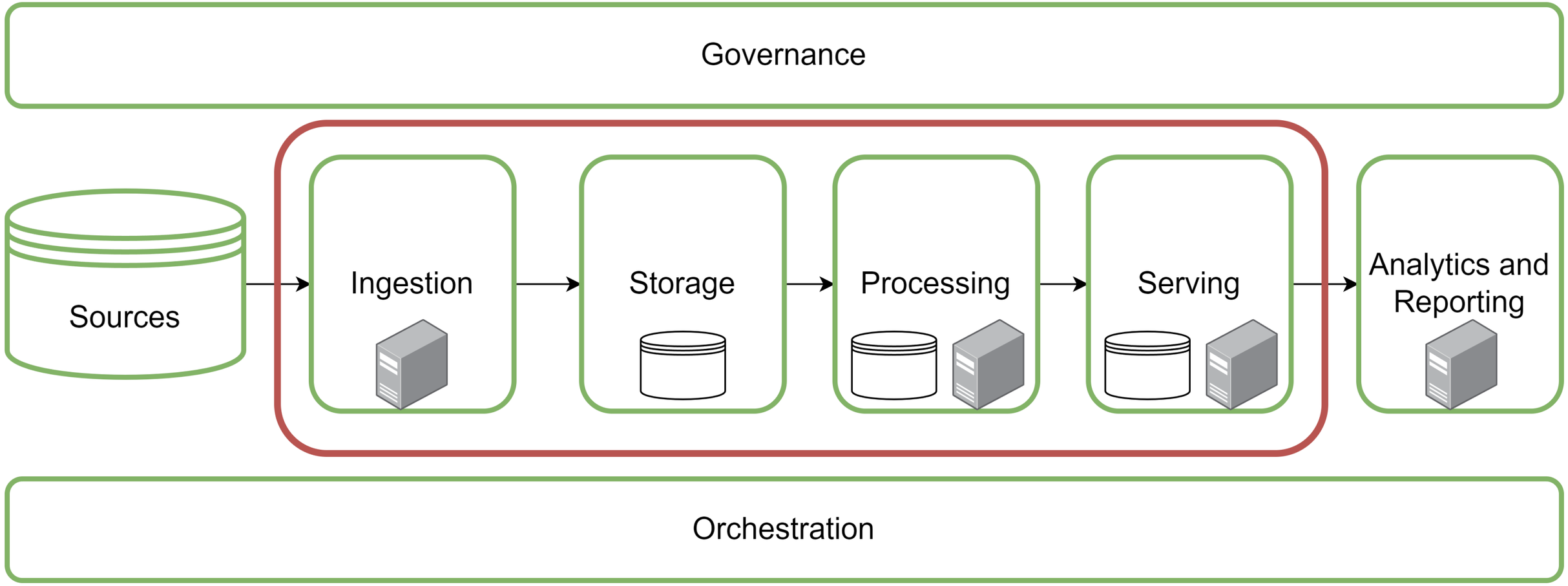




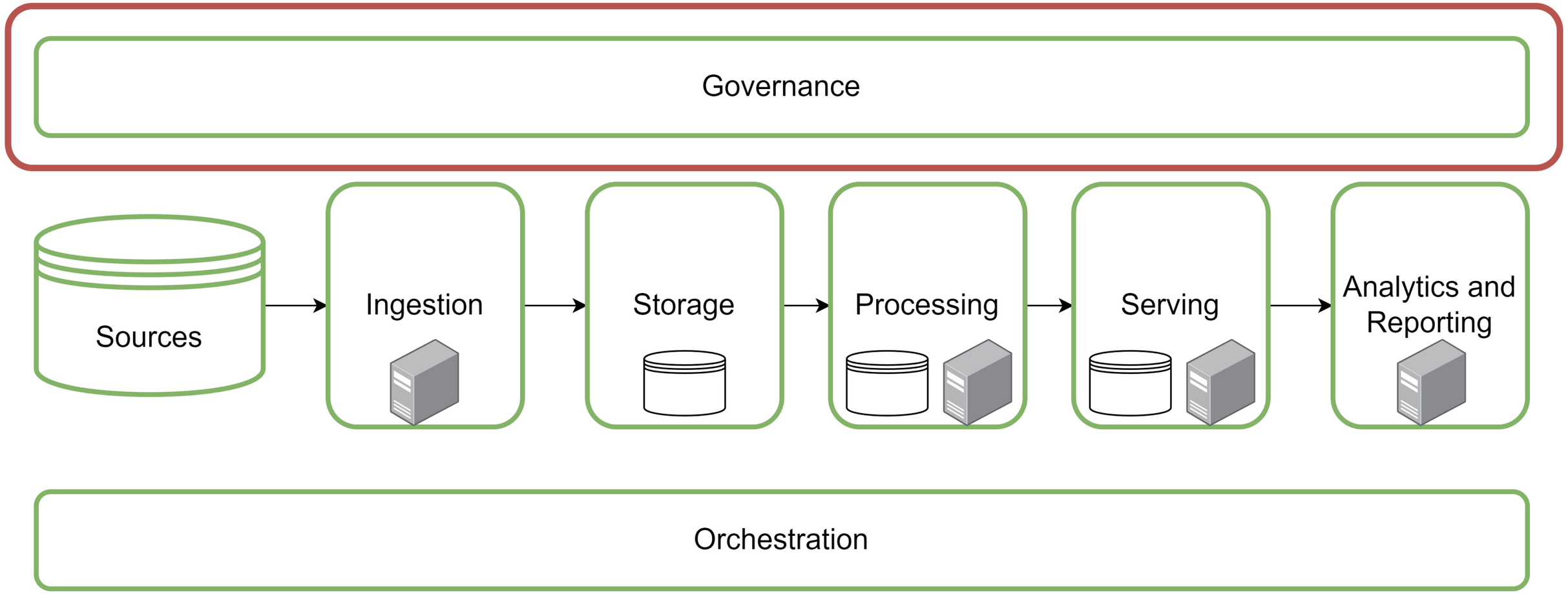


[1] https://martinfowler.com/articles/data-monolith-to-mesh.html

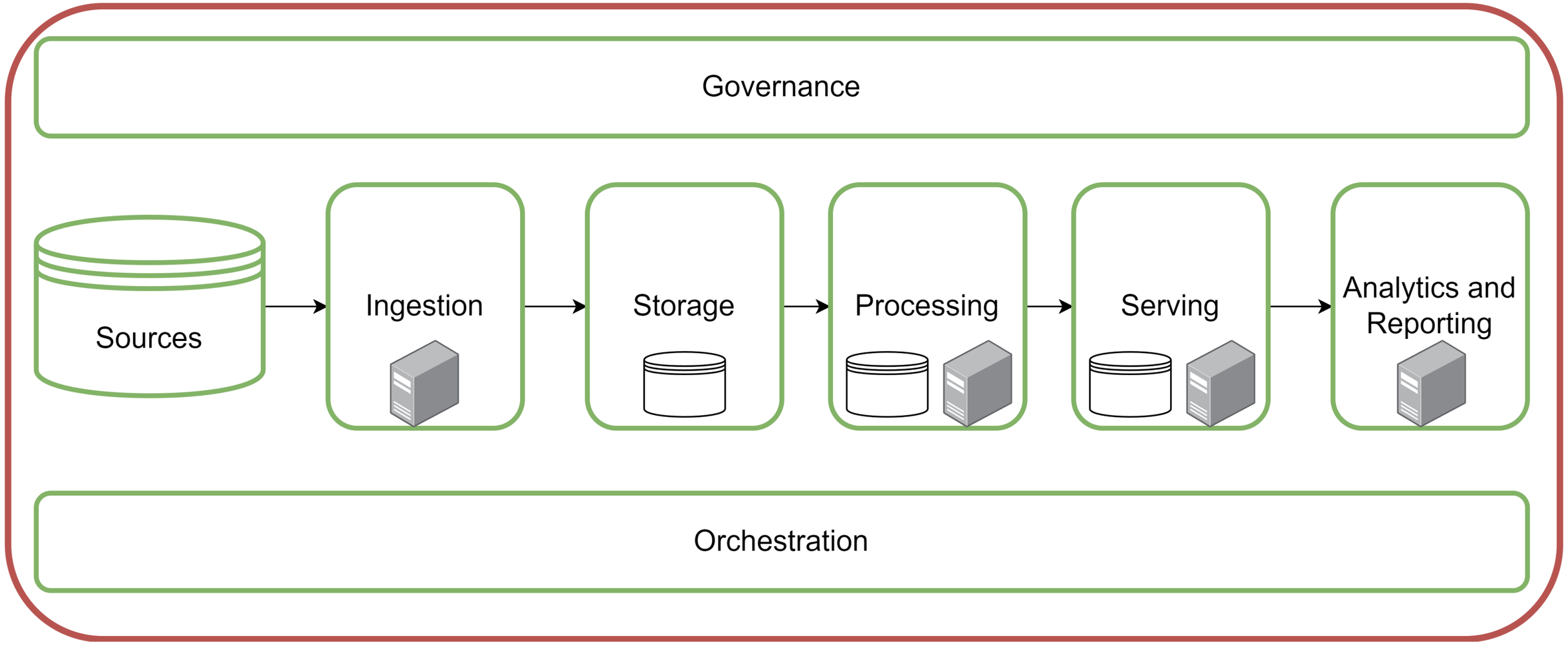
Chapter 2 - Modern Data Architecture Components



Chapter 3 - Transversal Components of Data Architectures



Chapter 4 - Putting it All Together



Congratulations!

UNDERSTANDING MODERN DATA ARCHITECTURE