Data Storage: Storage types

UNDERSTANDING MODERN DATA ARCHITECTURE



Miller Trujillo
Senior Software Engineer

datacamp

Blob storage

- BLOB = Binary Large Object
- Support all types of data, including non-structured
- Scalable. Virtually unlimited
- Blob storages are cheap!







Blob storage use cases

- Unstructured data - Media storage
- Backup and archiving
- Landing zone
- Content delivery

SQL

	SQL Databases
Use Cases	Transactional applications, applications with complex queries, data consistency and integrity
Types of Data Supported	Structured data (tabular)
Common Offerings	AWS RDS, GCP Cloud SQL, Azure SQL Database - MySQL, PostgreSQL, Data warehouse

NoSQL

	NoSQL Databases
Use Cases	High scalability and demand, eventual consistency
Types of Data Supported	Semi-structured data (JSON, XML, key-value, graph, etc), time series
Common Offerings	MongoDB, AWS DynamoDB, GCP Firestore, Azure Cosmos DB, Hbase

Data warehouses and data lakes

- BigQuery, Redshift, Snowflake, Databricks
- Blob storage (S3, Cloud Storage, Blob storage)
- Logical segregation
- ELT (Extract, load, transform)

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Data Ingestion

UNDERSTANDING MODERN DATA ARCHITECTURE

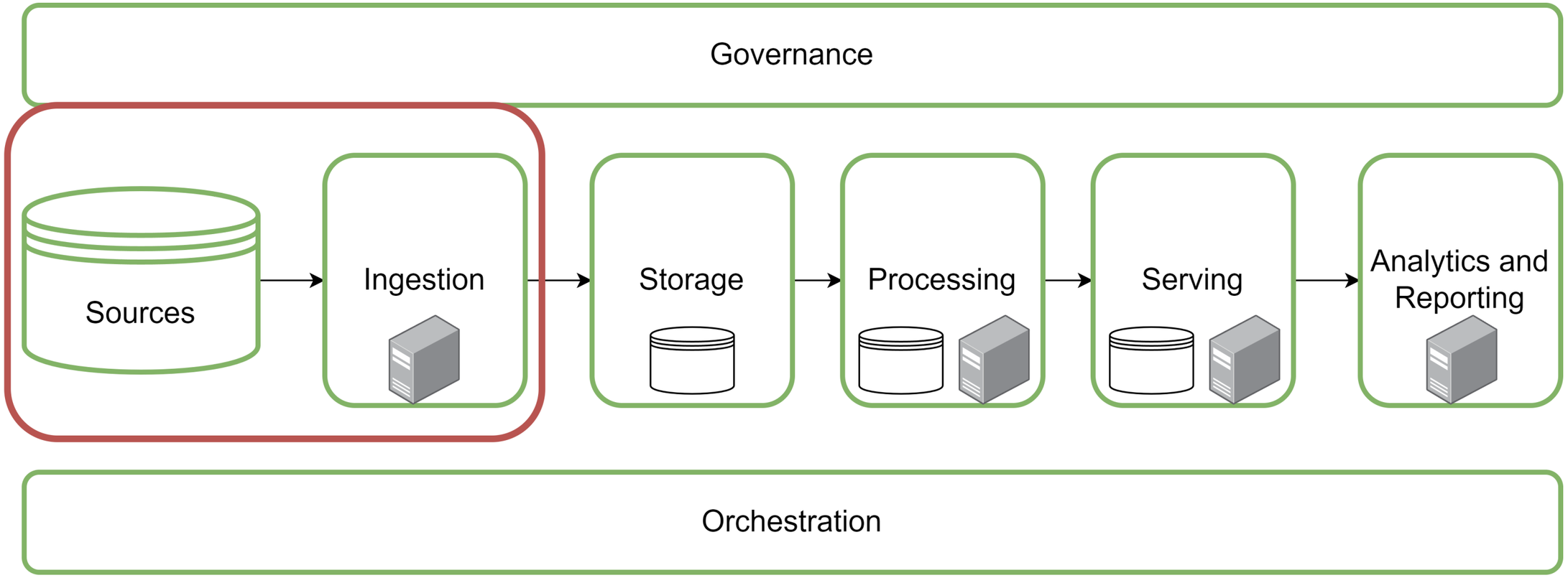


Miller Trujillo
Senior Software Engineer

datacamp

What is data ingestion?

- Functional requirements
- Functional can be impacted by analytics



Batch ingestion

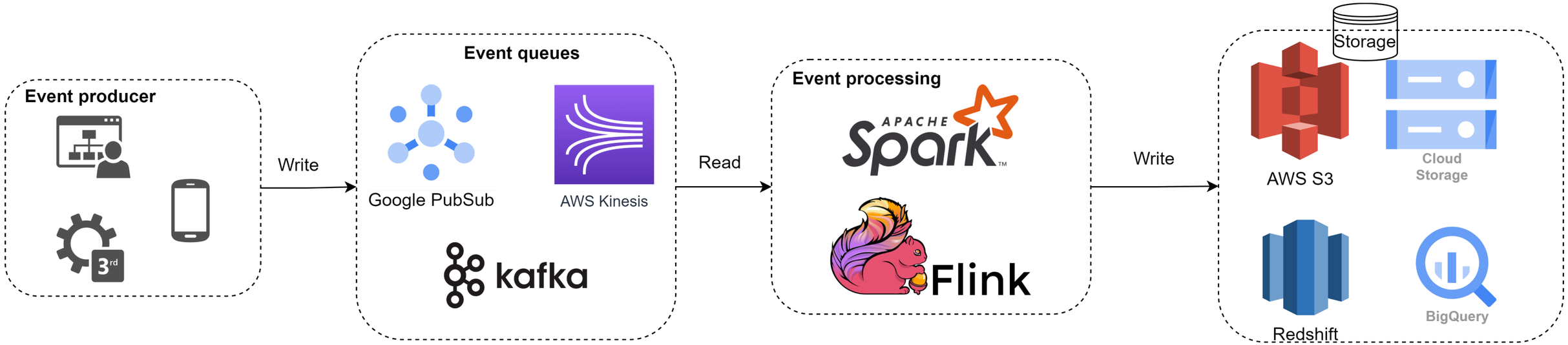
- Scheduled to ingest data periodically
- Copy in our platform for analytics
- Reading all data vs reading what is new to us
- Big datasets requires reading partially
- Smaller datasets could be overwritten

Batch ingestion: Bring only what changed

- Infinite resources are impossible
- Ingest only what has changed
- Updated at timestamp, or flag
- Latest state of data
- Deletion will require a flag or consolidation

Streaming ingestion

- Push model
- Event queues
- 24/7 compute
- Landing zone



Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Data Processing

UNDERSTANDING MODERN DATA ARCHITECTURE



Miller Trujillo
Senior Software Engineer

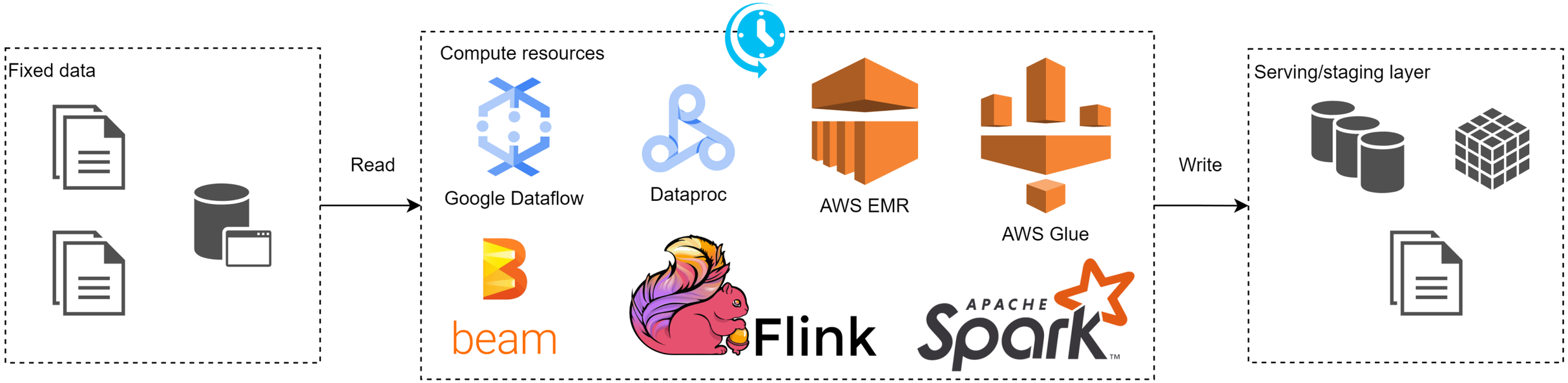
datacamp

What is data processing?

- Exploration
- Data quality: Checks and transformations
- Analysis
- Aggregations
- Transformations

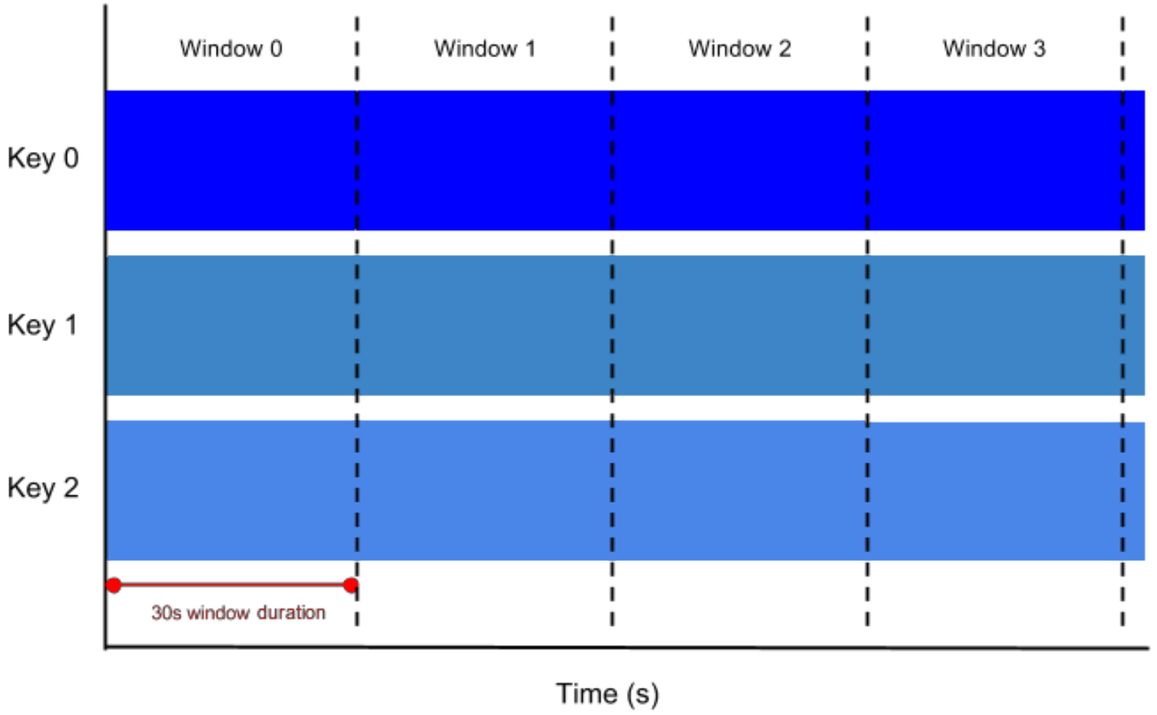
Batch processing

- Batch and streaming
- Fixed amount of data

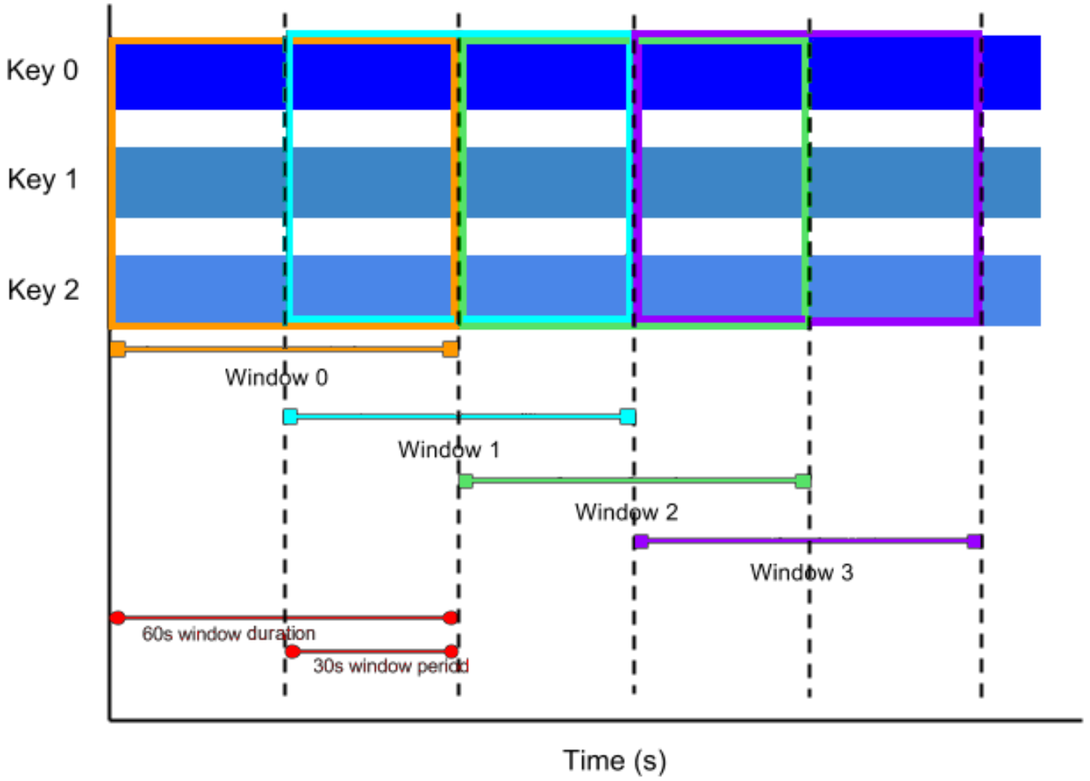


Streaming processing

Fixed time window



Sliding time window



[1] https://beam.apache.org/documentation/programming-guide/#windowing

Streaming processing concepts

- When data was generated
- When data arrived
- Watermarks
- Late data
- Trigger new processing

Processing technologies

Use case	Solution	Cloud solution
Batch/streaming, big data, cluster	Apache Spark, Flink, Beam	AWS EMR, AWS Glue, Google Dataproc, Google Dataflow
Batch/streaming, big data, serverless (servers are fully managed by the provider)	Apache Spark, Beam	AWS Glue, Google Dataflow
Individual events, simple processing, 24/7 support without servers running	General programming languages: Python, Javascript, C#, Java, Go	AWS Lambda, Google Cloud Functions

Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE

Data serving

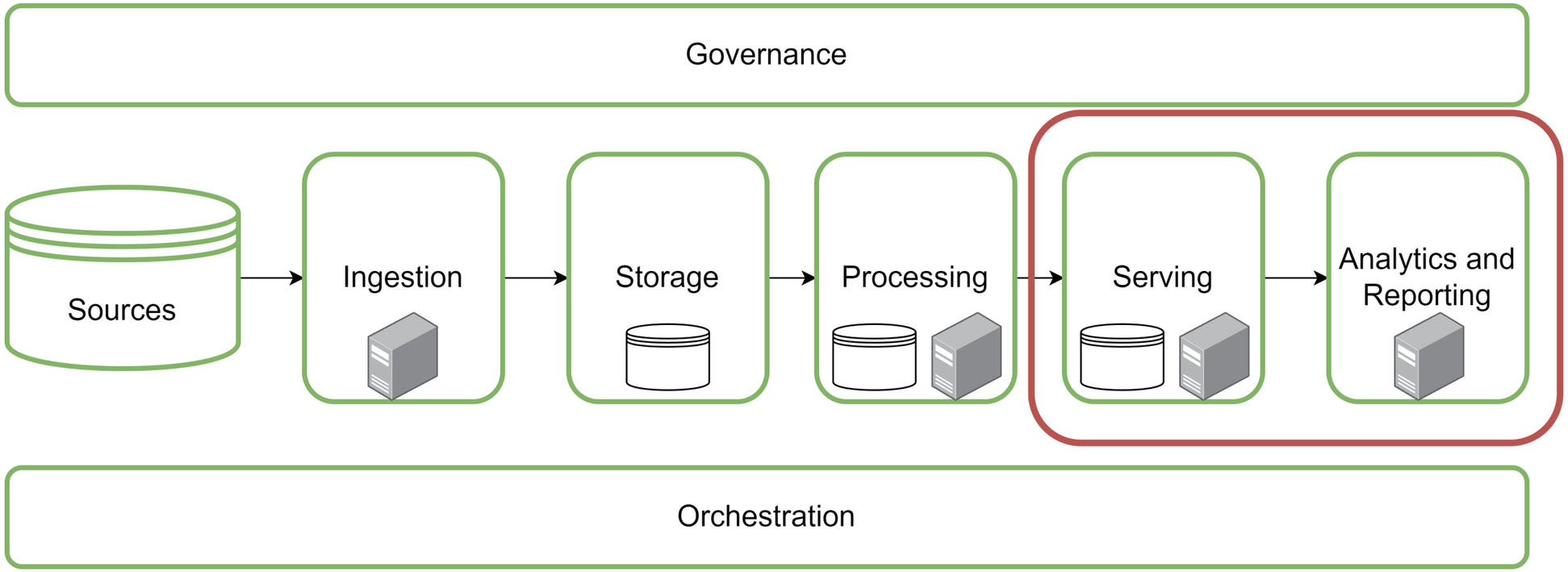
UNDERSTANDING MODERN DATA ARCHITECTURE



Miller Trujillo
Senior Software Engineer

datacamp

Serving layer



- Where to store processed data
- Protocols to consume it

What data do we have?

- What data do we have?
  - Is data structured?
  - Is our data a time series?
- How data will be consumed?








- Data Warehouse
  - Automated Reporting/Dashboarding
  - BI/Data Analytics
  - Direct Queries
- Blob storage
- Time series database
- NoSQL database

How will data be consumed?

- Data warehouse
  - Automated reporting/dashboarding
  - BI/Data analytics
  - Direct queries
- Machine learning models
- Applications
  - Summarized information
  - Individual records
- Not one size fits all!







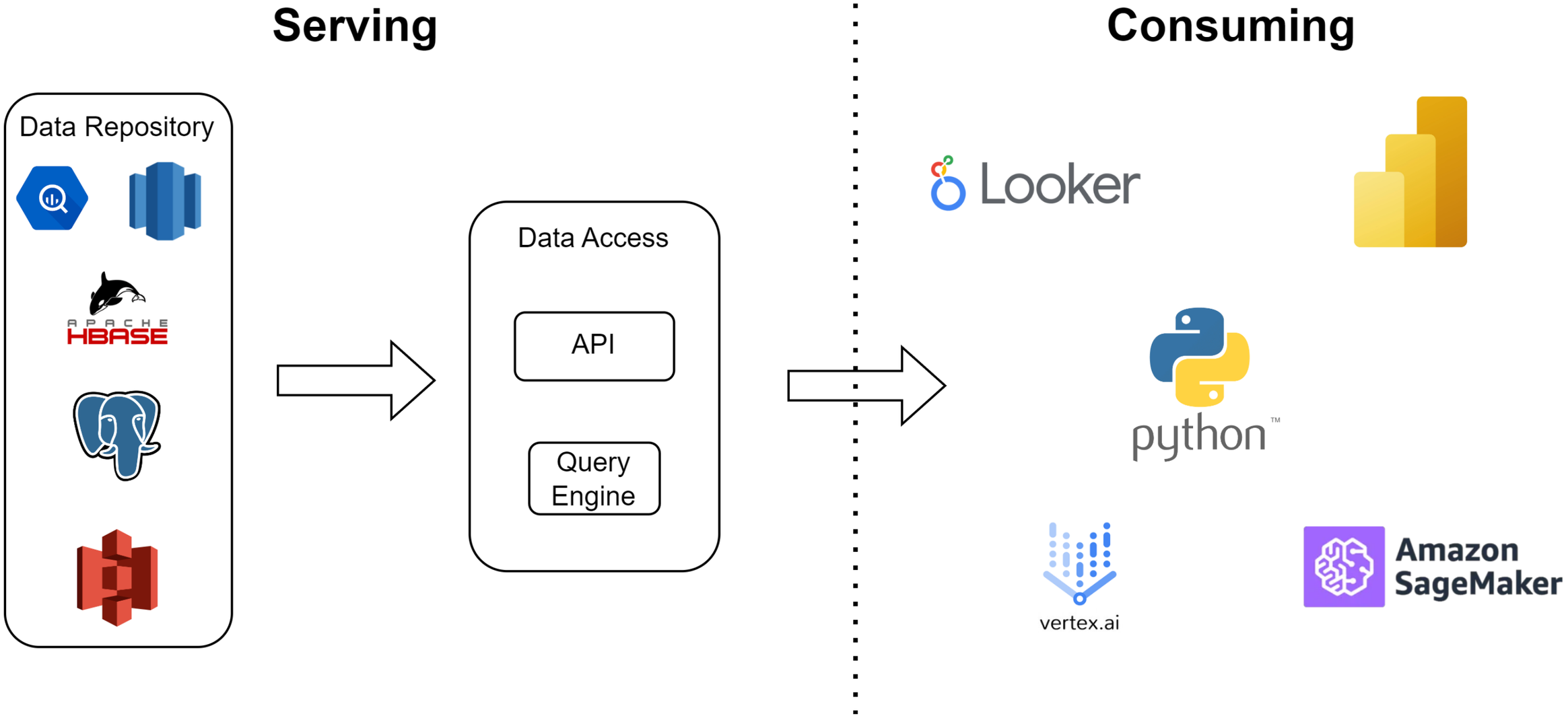
Amazon API Gateway



Serving your data depending on your use case

System	What type of data	Consuming use case	Tools
Data warehouse	Structured data, can handle semi-structured data	Analytical queries, BI/Analytics, reporting/dashboards	BigQuery, Redshift, Snowflake
Blob storage	All type of data	ML models with video/imaging processing, archiving	S3, Cloud Storage, Azure Storage
NoSQL database	Semi-structured data, time series	API exposure, high demand	BigTable, DataStore, DynamoDB, CosmosDB
RDMBS	Structured data, can handle semi-structured data	Individual records consumption, API exposure, lower demand	Postgres, MySQL, Oracle

Serving vs. Consuming



Let's practice!

UNDERSTANDING MODERN DATA ARCHITECTURE