dataiku
The Universal AI Platform™

Storage in the AI era

Arnaud Pichery

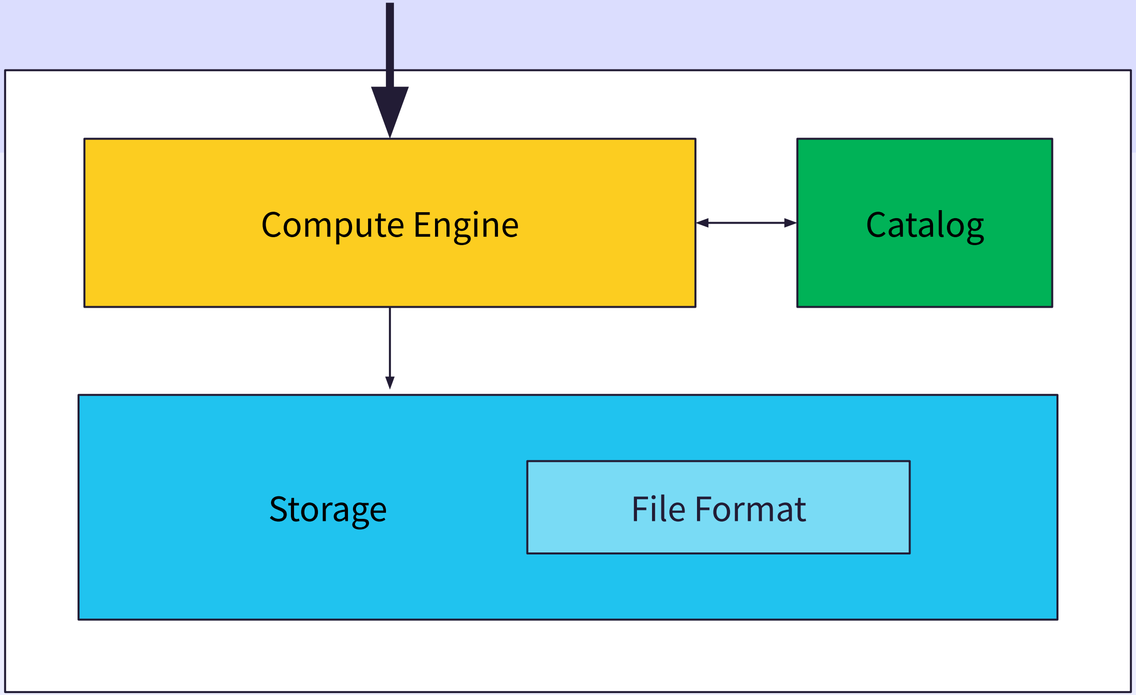
VP Engineering, Dataiku

01.

Data Warehouses

Data Warehouse

Analytics on Big Data



(Semi-)structured data

Separation of Storage and Compute

Massively Parallel Processing

Cloud Data Warehouses

SQL champions

SQL support

Cloud native architecture

Separation of Storage and Compute

Massively Parallel Processing (MPP)

Support for Modern Data Formats

snowflake

Google BigQuery

Azure Synapse Analytics

amazon REDSHIFT

databricks

Cloud Data Warehouses

SQL champions

⊕

Simplicity of use and administration

Decouples storage and compute

Infinitely scalable

Contains a storage engine that optimizes data layout

⊖

Data stored in proprietary format

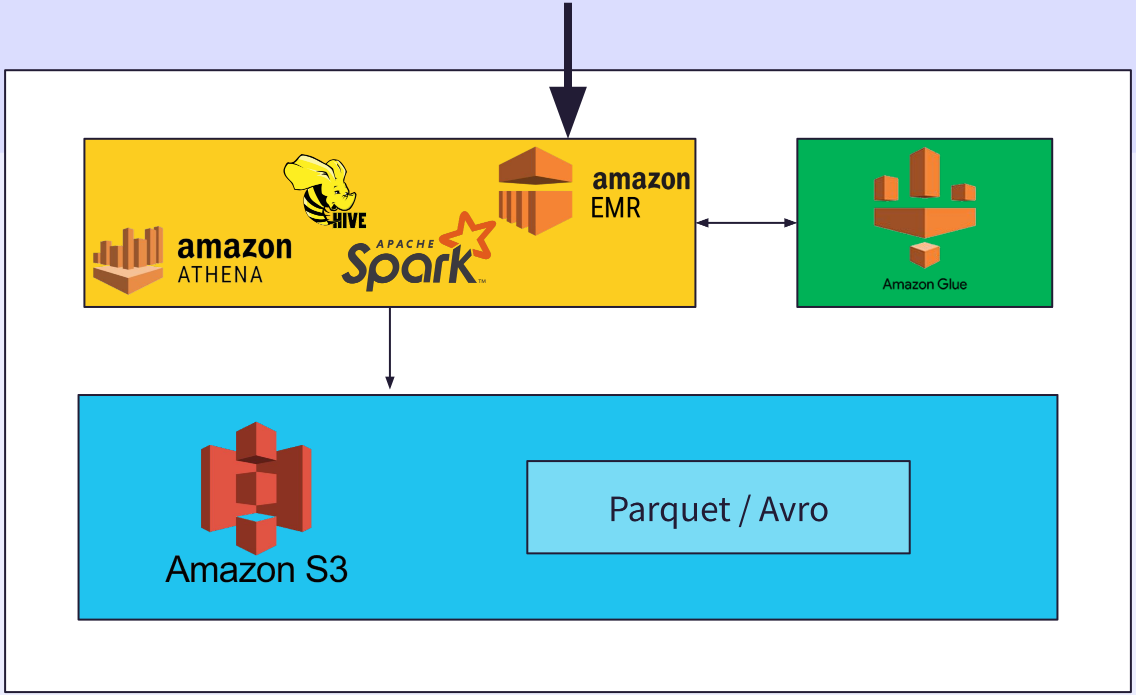
Need to land data in cloud storage to ingest

Expensive ($$$)

Cannot handle unstructured data nor non-SQL workloads

Data Lakes

Analytics on Big Data



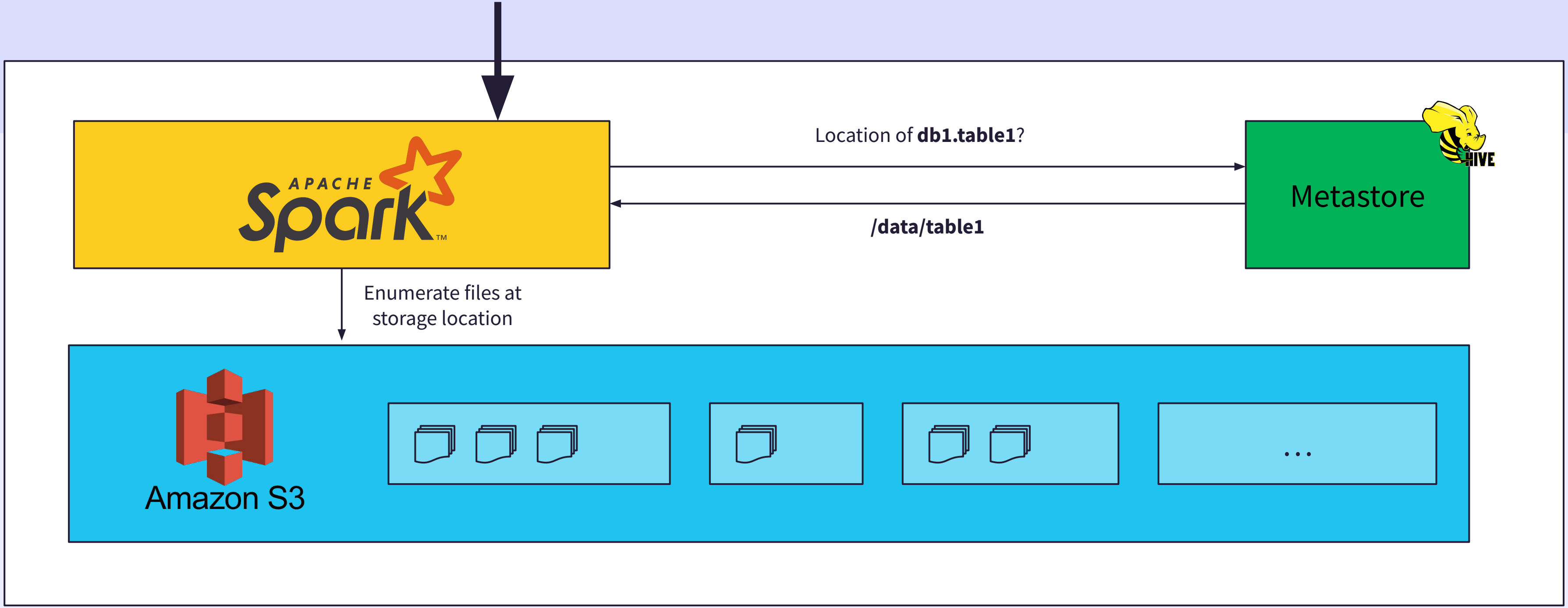
Hadoop-like architecture

Separation of Storage and Compute

Massively Parallel Processing

Data Lakes

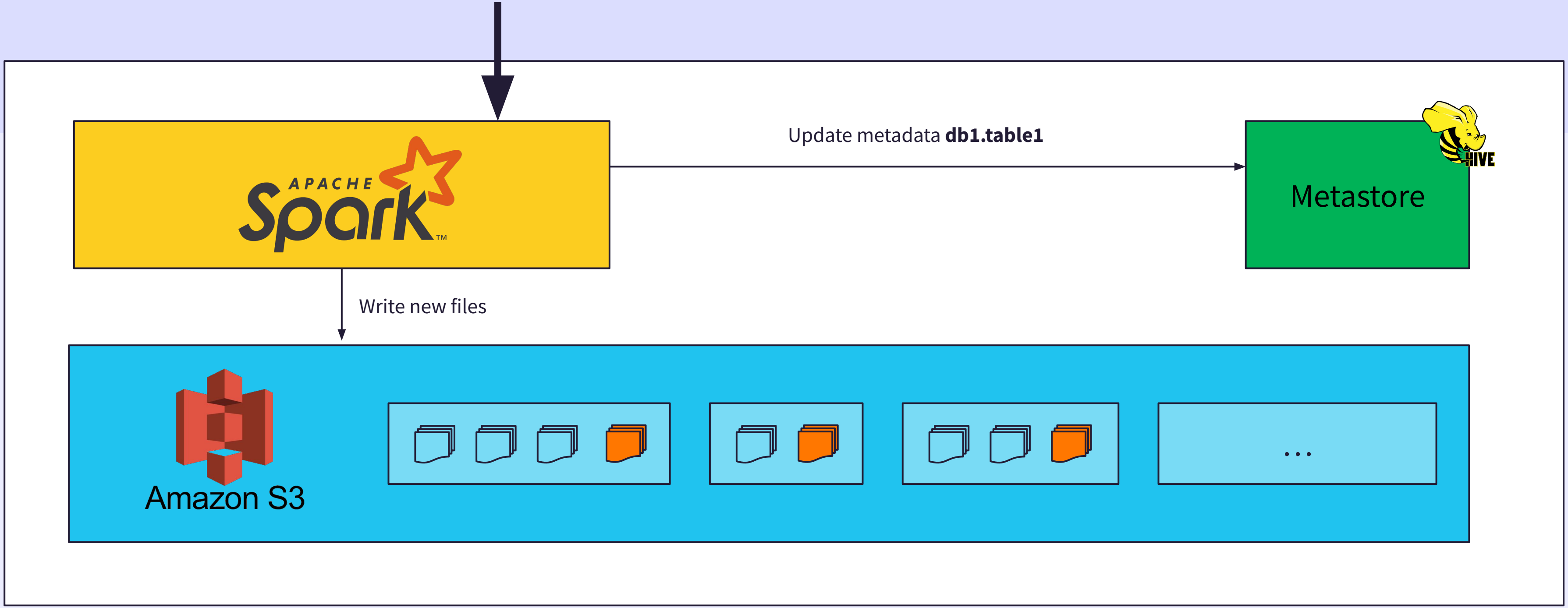
Analytics on Big Data



APACHE Spark
Location of **db1.table1**?
/data/table1
Metastore
HIVE
Enumerate files at storage location
Amazon S3
...

Data Lakes

Analytics on Big Data



Data Lakes

Analytics on Big Data

➕

Fully open-source storage formats

Modular architecture

Lower cost ($$)

Can process unstructured data

Provides SQL and Python/R/Scala APIs for data processing

⛔

Complex tuning of computation engines

Generally less performant than equivalent data warehouses

Lack of ACID transaction guarantees

The hot format

The new shiny thing

Data Lakes on Iceberg

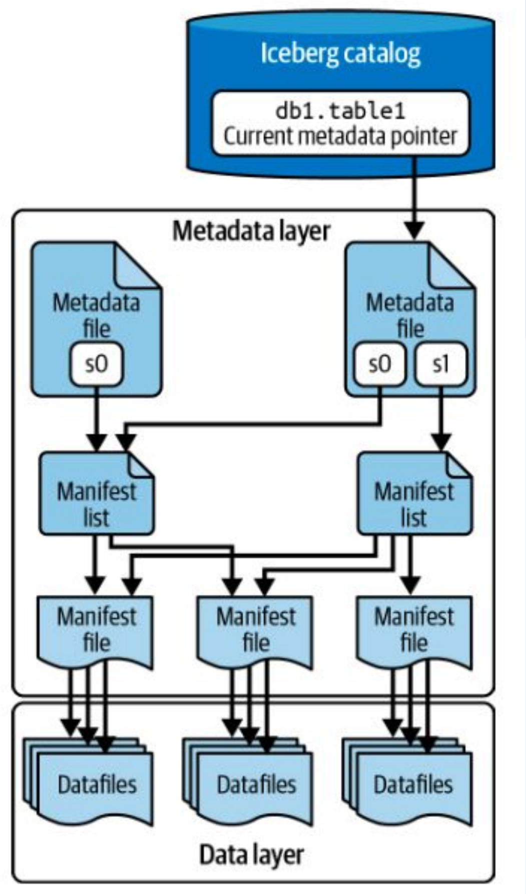
Open-source champion

Iceberg catalog points to location of a metadata file.

Metadata file contains references to one or more data snapshots, current snapshot, partition information and schema.

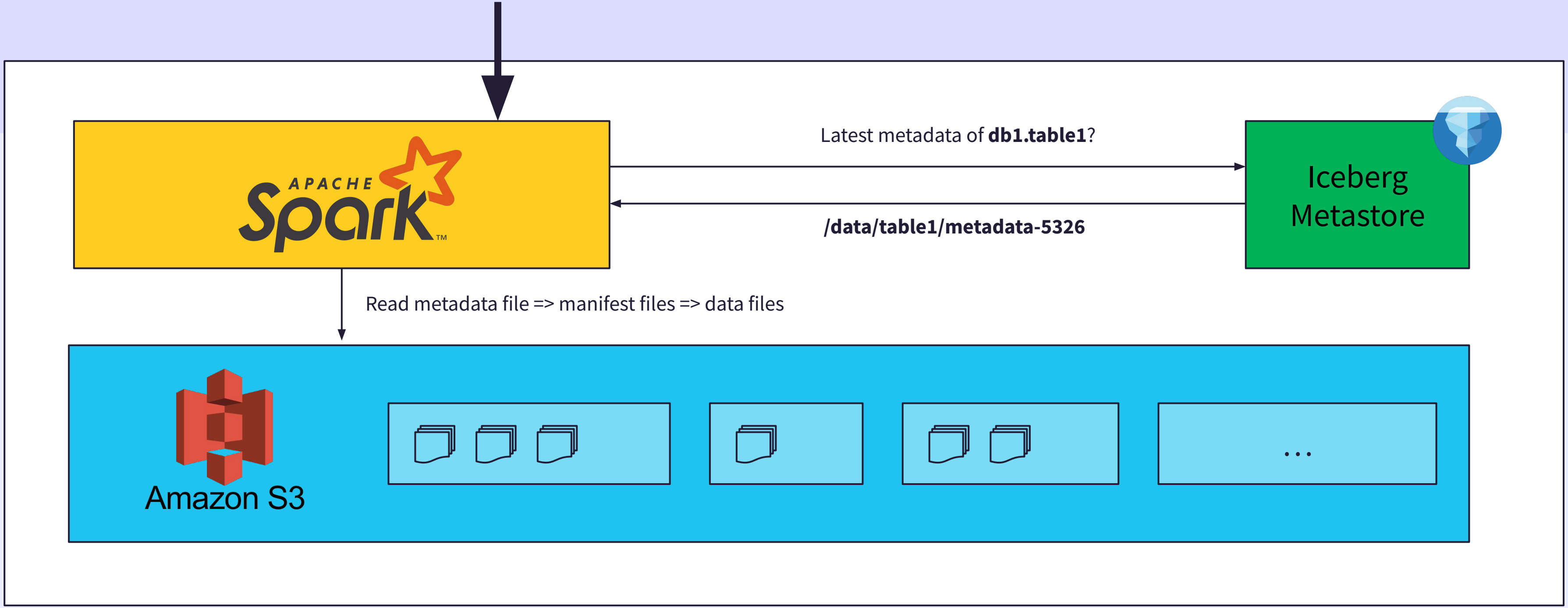
Manifest list is a snapshot of an Iceberg table at a point in time. Contains references to manifest files.

Manifest files keep track of datafiles and stats (e.g. partition membership, record counts, max/min, etc.).



Data Lakes on Ice(berg)

Upright spin



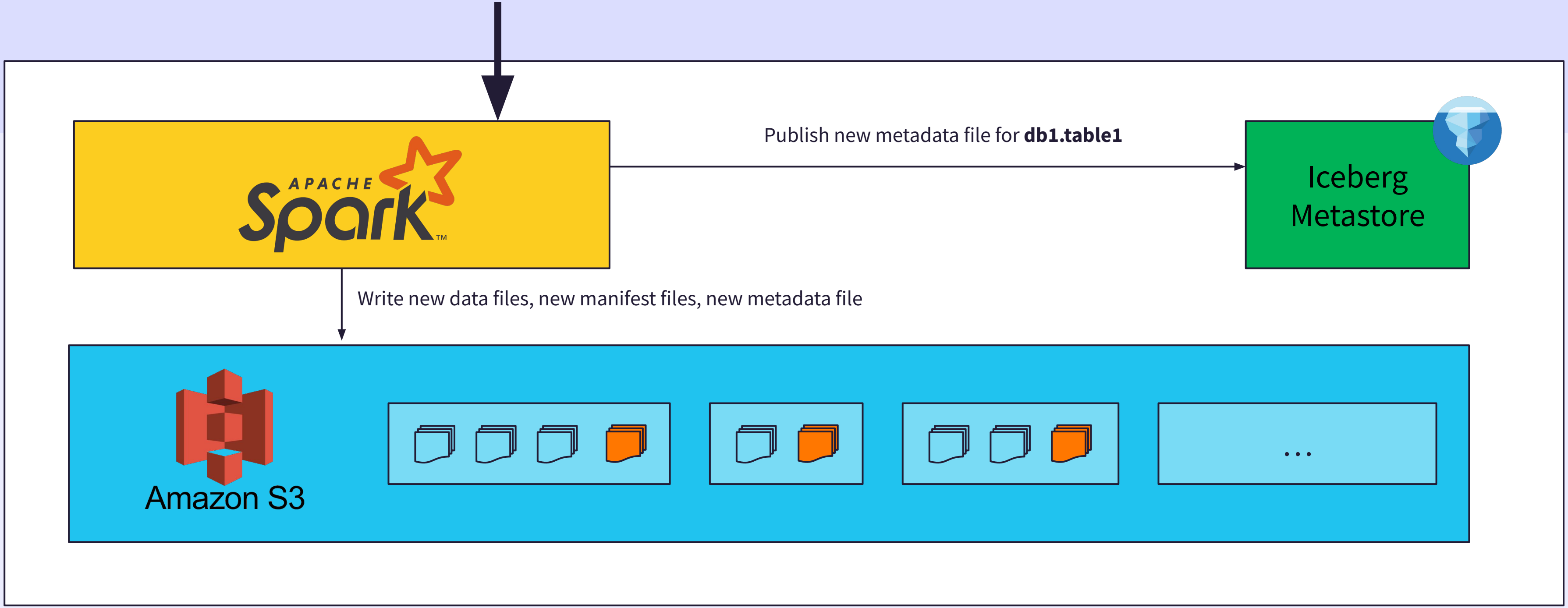
Latest metadata of **db1.table1**?

/data/table1/metadata-5326

Read metadata file => manifest files => data files

Data Lakes on Ice(berg)

Triple axel jump



Publish new metadata file for **db1.table1**

Write new data files, new manifest files, new metadata file

02.

Vector Stores

Vector databases

New kids on the block

RAG (Retrieval-Augmented Generation)

- → Store billions of vectors
- → Hybrid storage
    - ○ Unique ID
    - ○ Vector (384/512 doubles)
    - ○ Metadata (tags, timestamp, …)
    - ○ Raw data (text, image, pdf, ...)
- → Fast lookup of similar vectors

milvus

Pinecone

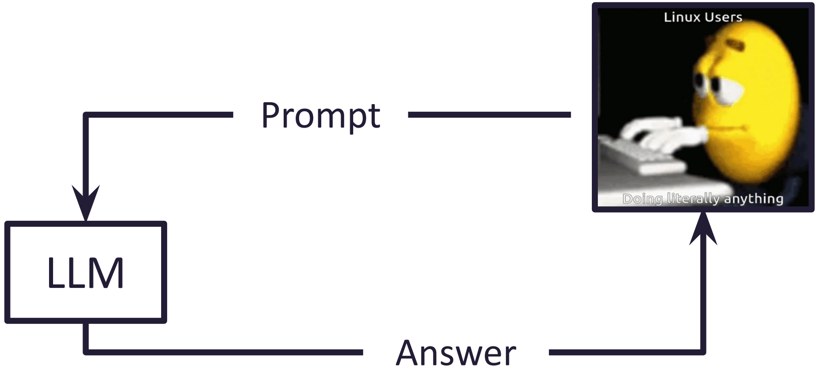
Chroma

Weaviate

Vector databases

New kids on the block

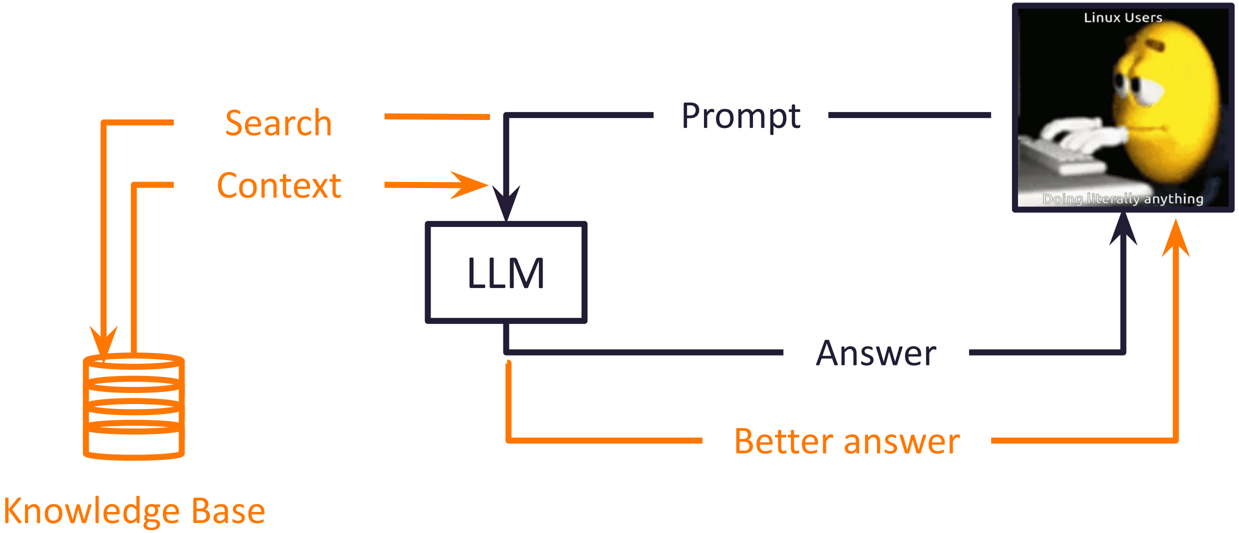
RAG



Vector databases

New kids on the block

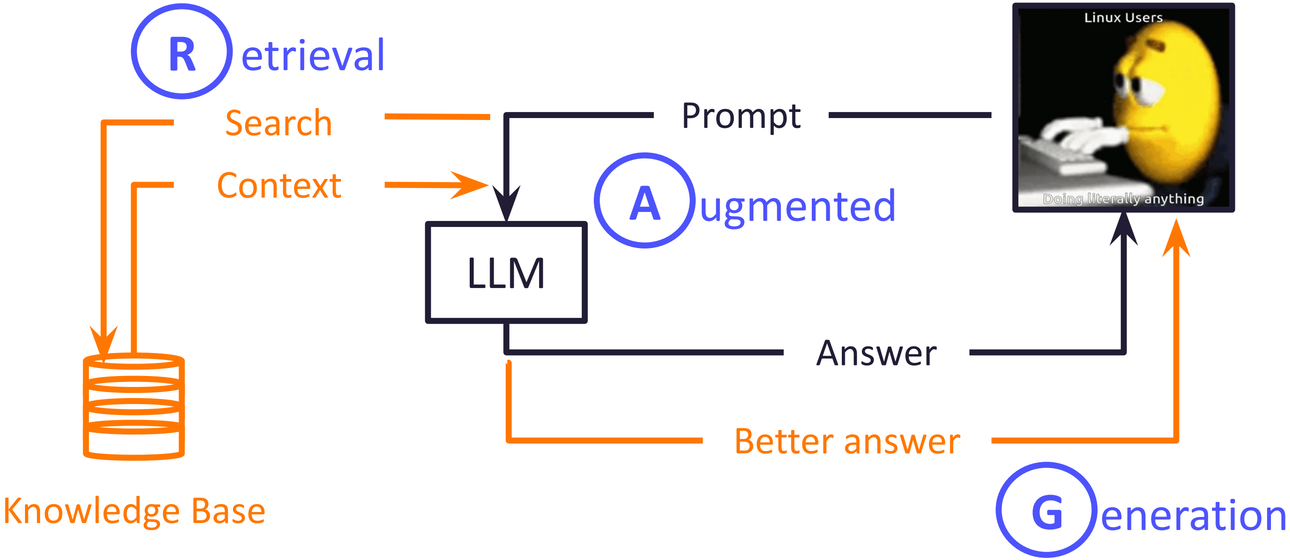
RAG



Vector databases

New kids on the block

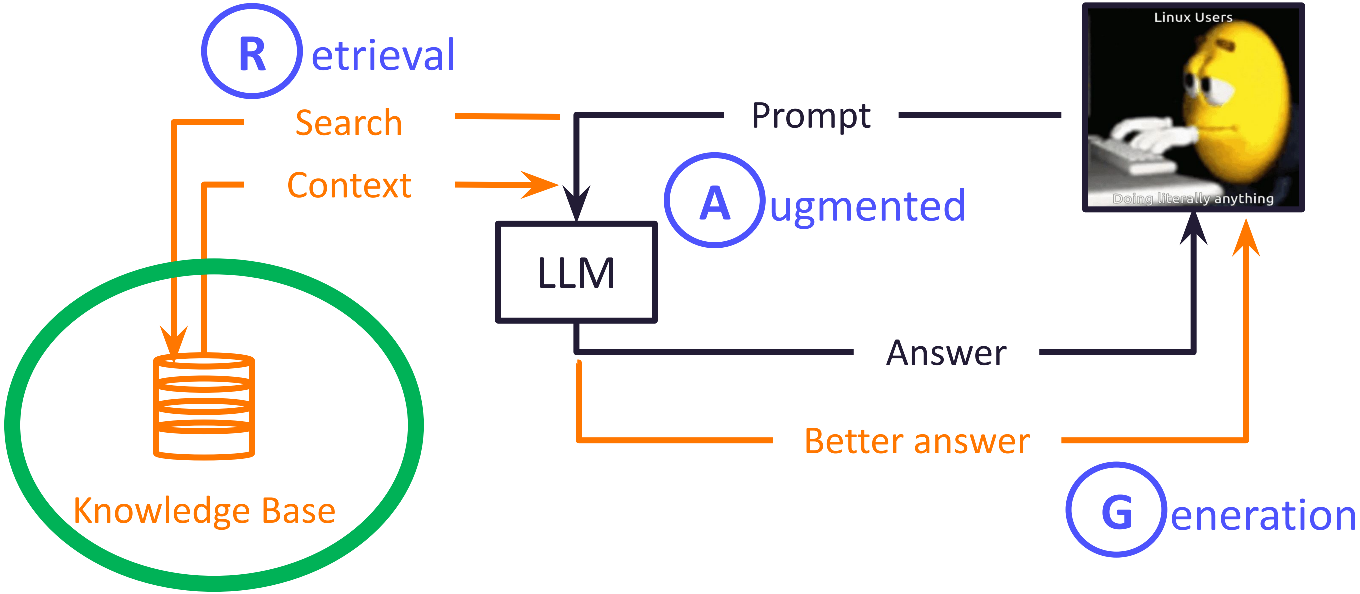
RAG



Vector databases

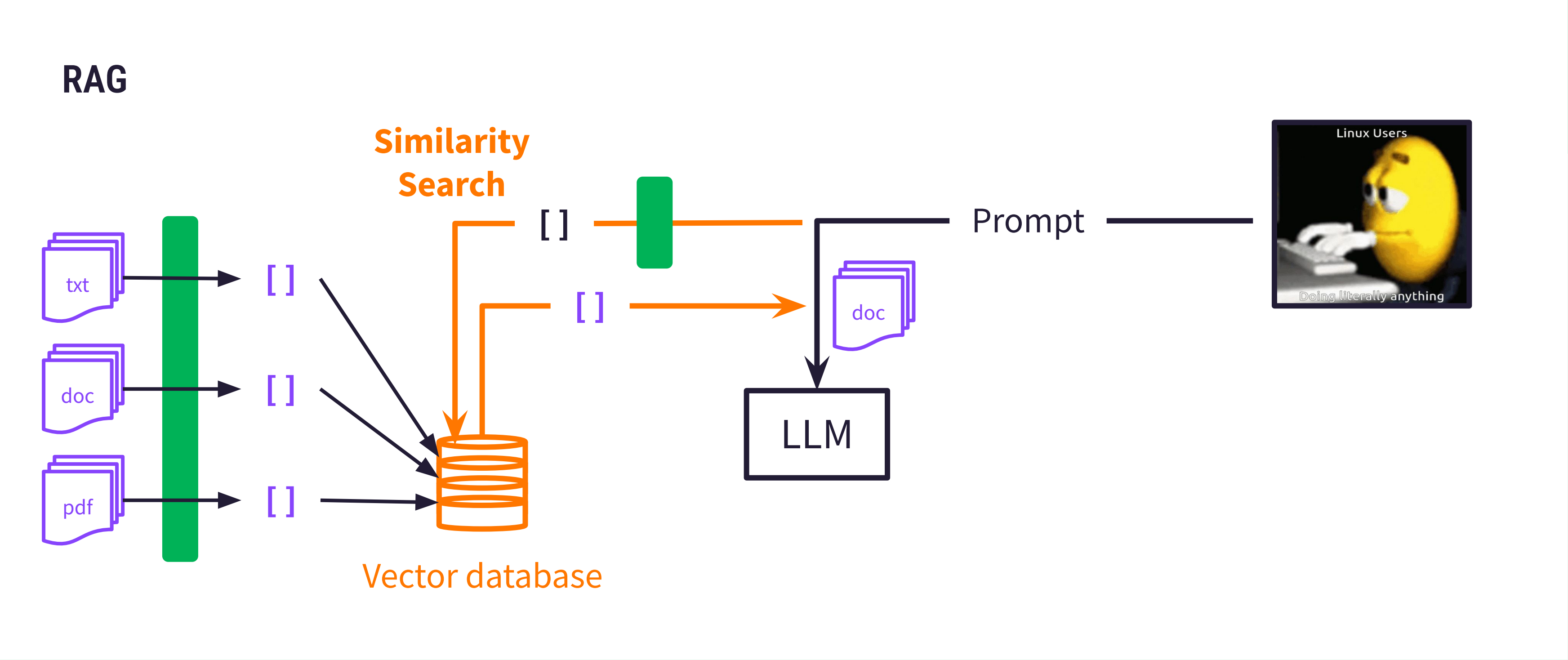
New kids on the block

RAG



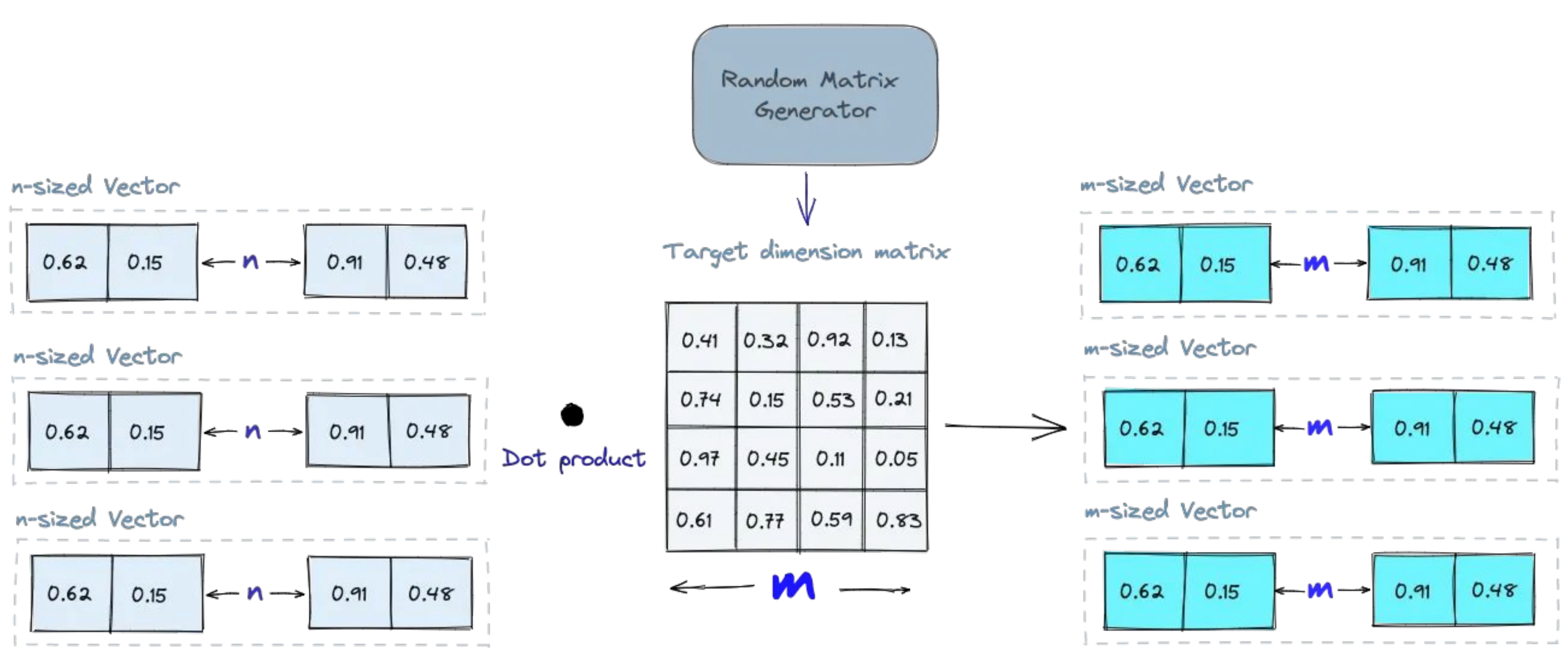
Vector databases

New kids on the block



Vector databases

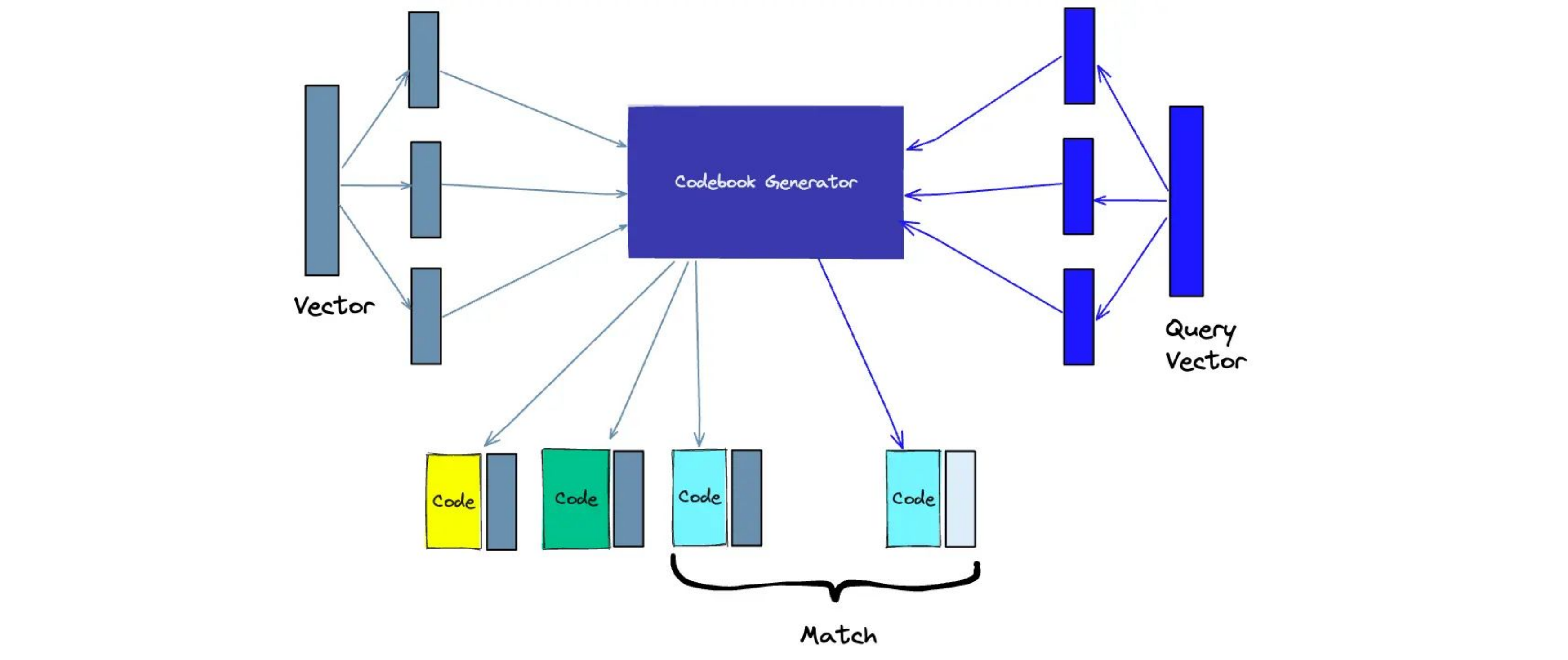
Random Projection



0.41	0.32	0.92	0.13
0.74	0.15	0.53	0.21
0.97	0.45	0.11	0.05
0.61	0.77	0.59	0.83

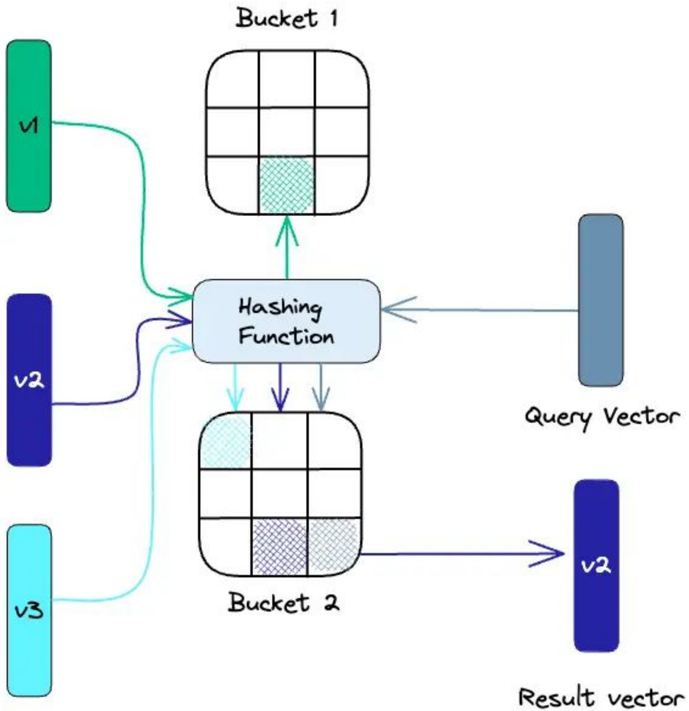
Vector databases

Product Quantization



Vector databases

Locality-sensitive hashing



dataiku
The Universal AI Platform™

Thank You

Do you have any questions?

dataiku

Meet Dataiku

The Universal AI Platform™