DATA ENGINEERING FOR LLMS

BY ABDULQUADRI OSHOARE

OUTLINE

1. Data Engineering
2. Traditional Data Engineering
3. Data Engineering 2.0
4. LLMs
5. RAGS
6. LTE
7. Vector Databases
8. Data Engineering for LLM
9. Challenges
10. Questions

DATA ENGINEERING

“Data Engineering involves designing and building systems for collecting, storing, and analysing data at scale.”

TRADITIONAL DATA ENGINEERING

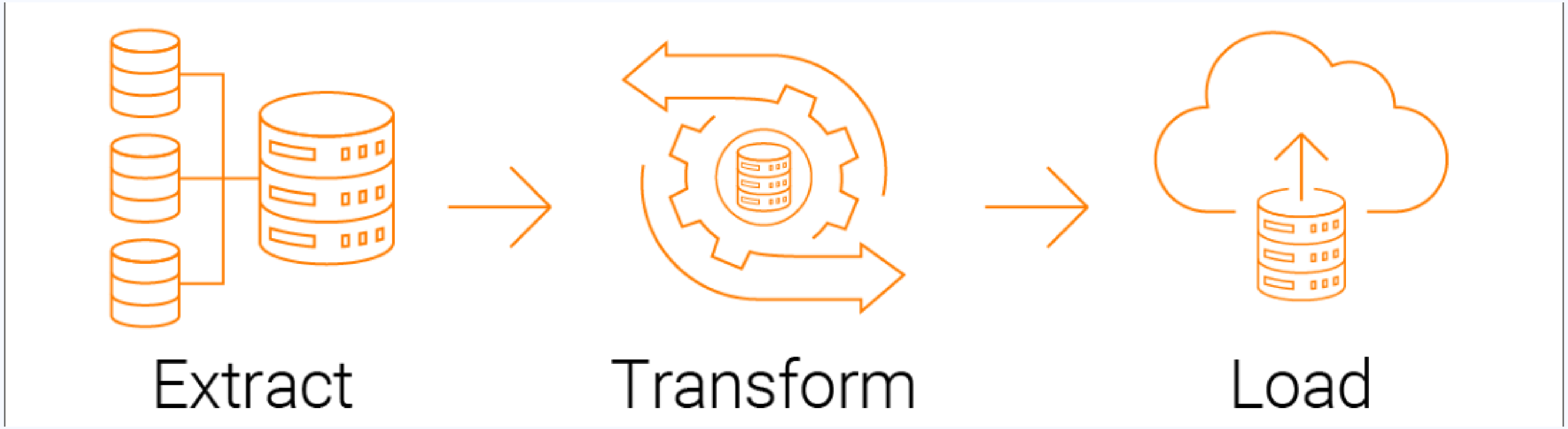
Data Engineers initially focused on creating and maintaining data infrastructures to ensure data was accessible for human users.

As the need for advanced analytics and processing large data volumes grew, their objective became not just building the architecture, but also structuring data to be quickly accessible, interpretable, and valuable to analysts, data scientists, and decision-makers.

KEY COMPONENTS

- Ingestion
- Transformation
- Storage
- Quality
- Governance
- Security

TRADITIONAL DATA ENGINEERING FLOW



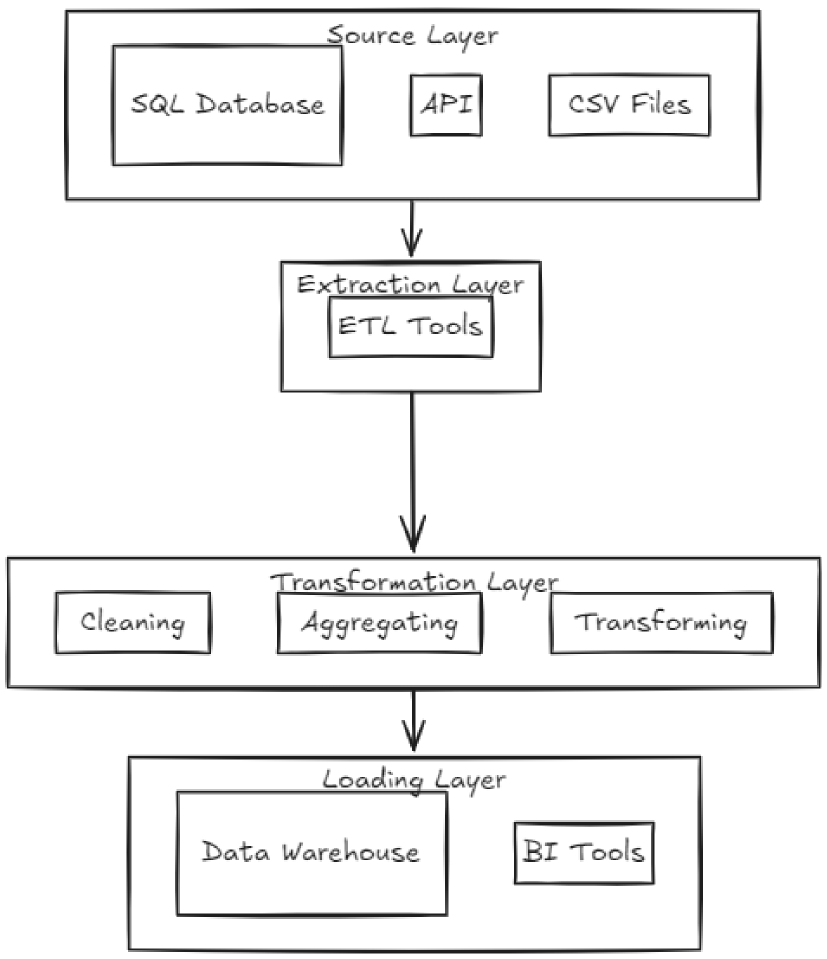
ETL (EXTRACT TRANSFORM LOAD) PROCESS

Extract: Data is pulled from various sources like databases, APIs, and files.

Transform: The extracted data is cleaned, aggregated, and transformed into a useful format.

Load: The transformed data is then loaded into a destination like a data warehouse or analytics tool for further analysis and reporting.

Types
Batch processing & Real-time processing



ETL Data Engineering Process

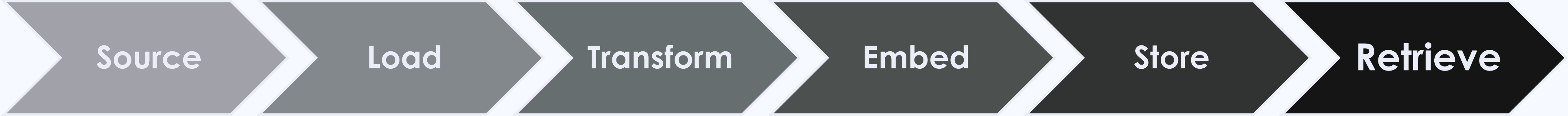
DATA ENGINEERING TOOLS

INGESTION	TRANSFORMATION	STORAGE	DATA QUALITY	WORKFLOW ORCHESTRATION
Airbyte	Dbt	Cloud Storage	Great Expectations	Apache Airflow
Apache Kafka	Apache Spark	BigQuery	Deequ	Prefect
Apache Nifi	Airflow	AWS S3	Tecton	Dagster
Databricks	Fivetran	Postgres, MySQL, MSSQL	Monte Carlo	

DATA ENGINEERING 2.0

“Data Engineering 2.0 involves providing data that is not only understandable for humans, but also for AI systems (LLMs).”

ETL & LTE



Source	Load	Transform	Embed	Store	Retrieve
Source systems, e.g. database, pdf, api e.t.c	Extracting data from source	Splitting, chunking of data	Convert to vectors	Store vectors in vector Db for easy access	Data is retrieved by LLMs

LLMS (LARGE LANGUAGE MODELS)

Large Language Models, are advanced AI models that can understand and generate human language. They are trained on massive datasets to predict and generate text.

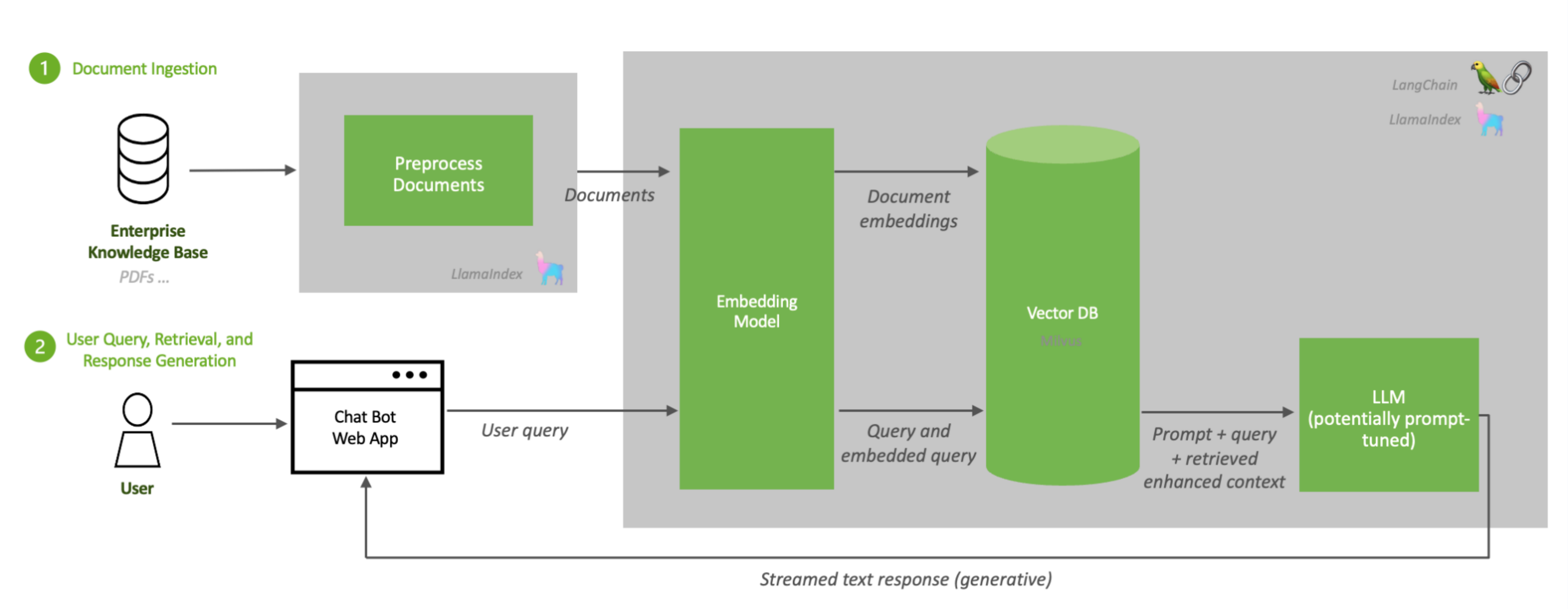
Examples:

GPT-4
BERT
Claude
Llama

RAGS

Retrieval-Augmented Generation is a method that enhances LLMs by combining them with external knowledge sources, enabling the generation of more accurate and contextually relevant information.

RETRIEVAL AUGMENTED GENERATION STEPS



Copyright nvidia

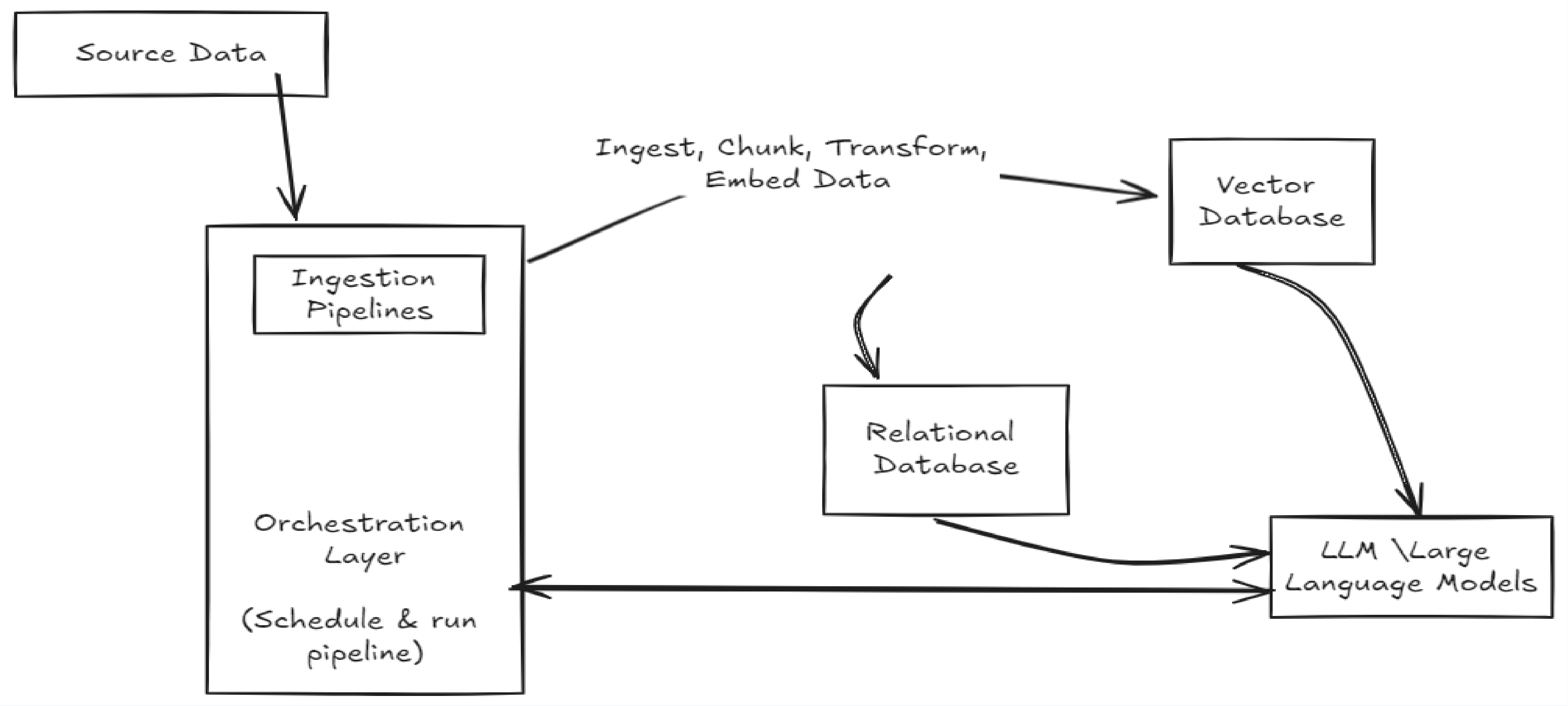
VECTOR DATABASES

A vector database is a collection of data stored as mathematical representations. Vector databases make it easier for machine learning models to remember previous inputs, allowing machine learning to be used to power search, recommendations, and text generation use-cases.

COMMON VECTOR DATABASES

Weaviate
milvus
qdrant
Pinecone
SingleStore
Chroma
zilliz

DATA ENGINEERING FOR LLM


Source Data
Ingest, Chunk, Transform, Embed Data
Vector Database
Ingestion Pipelines
Relational Database
Orchestration Layer
(Schedule & run pipeline)
LLM \Large Language Models

CHALLENGES

SCALABILITY

MODEL BIAS

OBSERVABILITY

DATA PRIVACY

DATA QUALITY

Questions ?

THANK YOU!