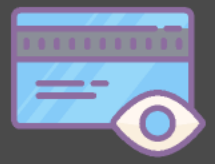
Use case examples utilizing streaming data

DATASIENS

- **Streaming data refers to continuous and real-time flow of data from various sources.**
- **Unlike batch data processing, where data is collected and processed in batches, streaming data processing involves the processing of data as it arrives, allowing for real-time analysis, decision-making, and action.**

Finance



Payment Fraud

Detecting fraudulent transactions in online transactions.

Manufacturing

Predictive maintenance

Predictive maintenance involves using machine learning models to predict when equipment or machinery is likely to fail

Retail

Recommender

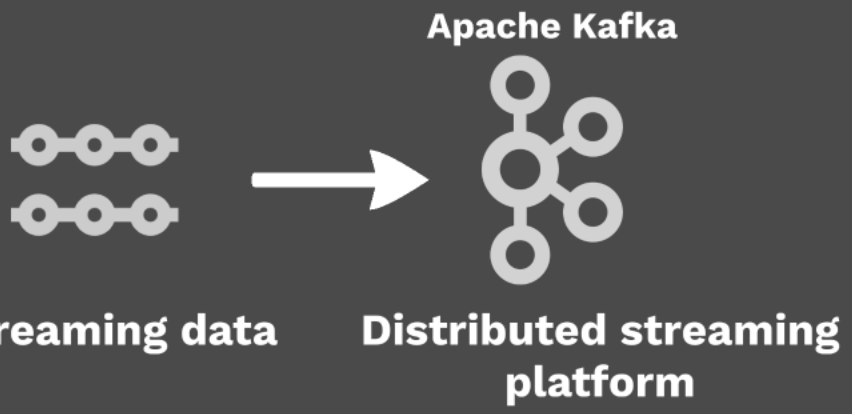
Retail recommender systems use algorithms to analyze customer data, such as purchase history, browsing behavior, and product preferences, to make personalized recommendations to customers.

Kafka 101

DATASIENS

Apache Kafka is an open source software for implementing a **distributed platform** for hosting **high througput streaming data**.

Commericial support is provided by Confluent.



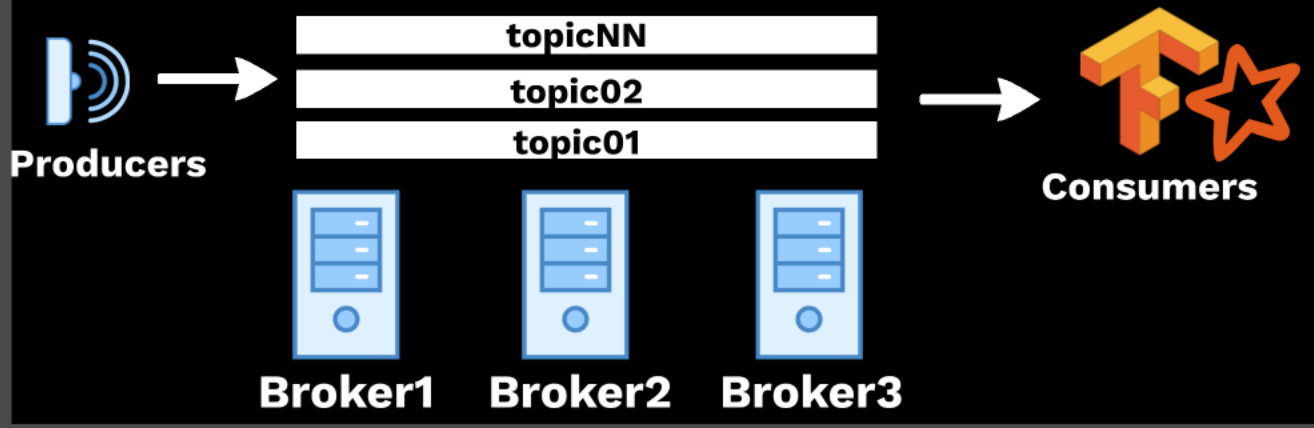
Kafka Glossary

Producer: Source of event data like sensors, log files etc.

Topics: The producers write messages to named topics. Like sensor01 will write messages to topic named topic-sensor01.

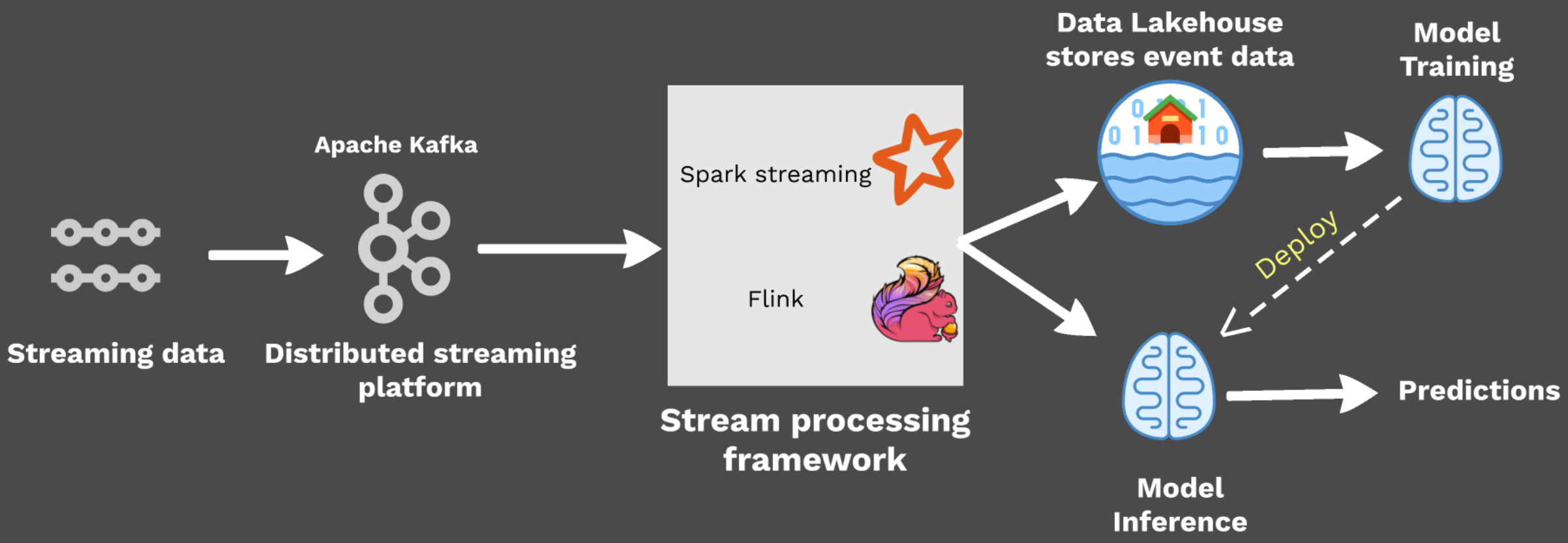
Broker: The server that hosts the topics. There are usually more than three brokers for fault tolerance.

Consumer: Applications that read/consume the messages. Like Tensorflow IO reading messages into a dataframe for training models or a spark application reading and processing messages from the topics.

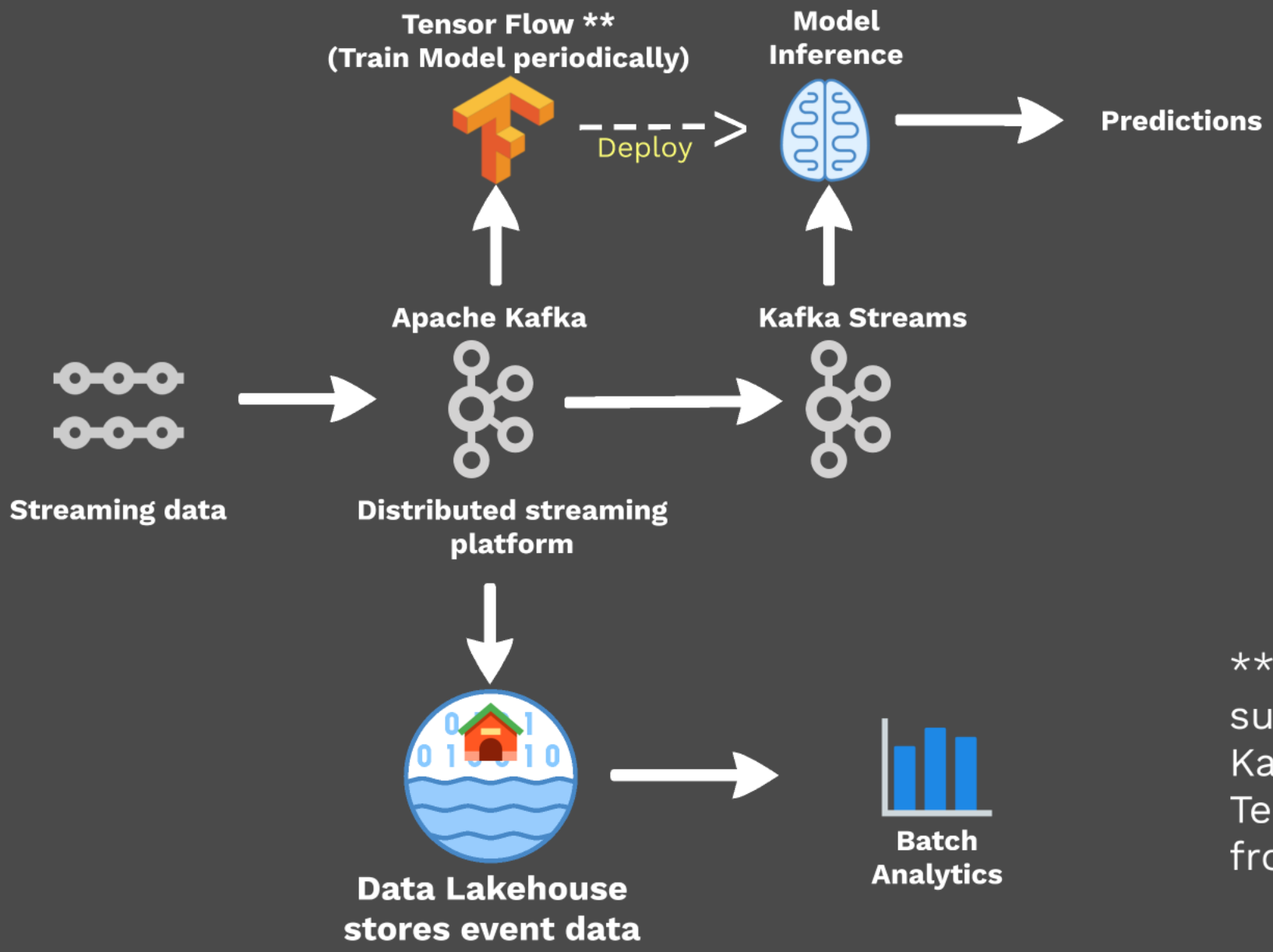


Lambda architecture for streaming data

DATASIENS



Kappa architecture for streaming data

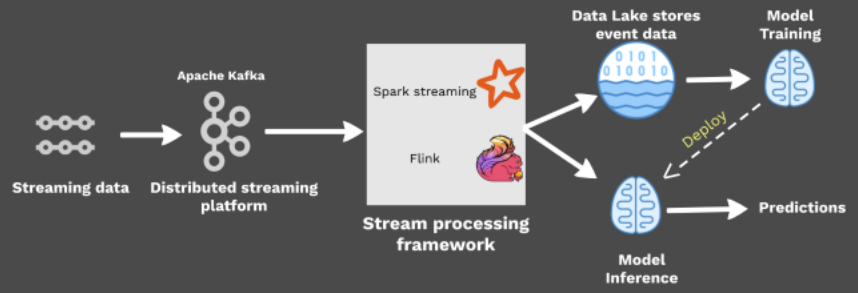


** Not all ML frameworks support native integration with Kafka. Tensorflow provides Tensorflow IO which can read from Kafka topics directly.

Lambda versus Kappa

DATASIENS

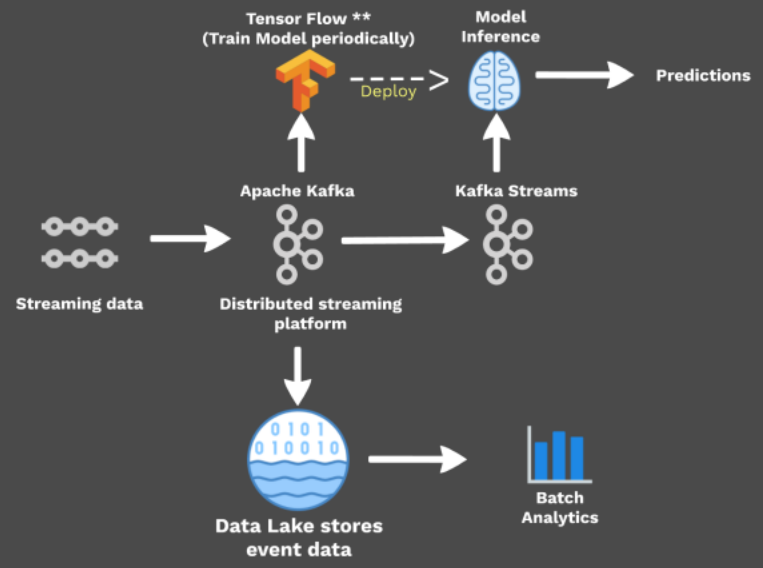
Lambda



Pros
- More prevalent and older architecture
- Easier to implement due to availability of know how.
- All model frameworks can be incorporated.

Cons
- Seperate pipelines/ technology for batch and real time.
- Model training infrequent due to time taken to offload data from Kafka to data lake.

Kappa



Pros
- Same pipeline/technology used for batch and streaming.
- Model training more frequent due to direct integration with Kafka.

Cons
- Only Tensorflow natively supports training models directly from Kafka.
- Implementation know-how still scarce.