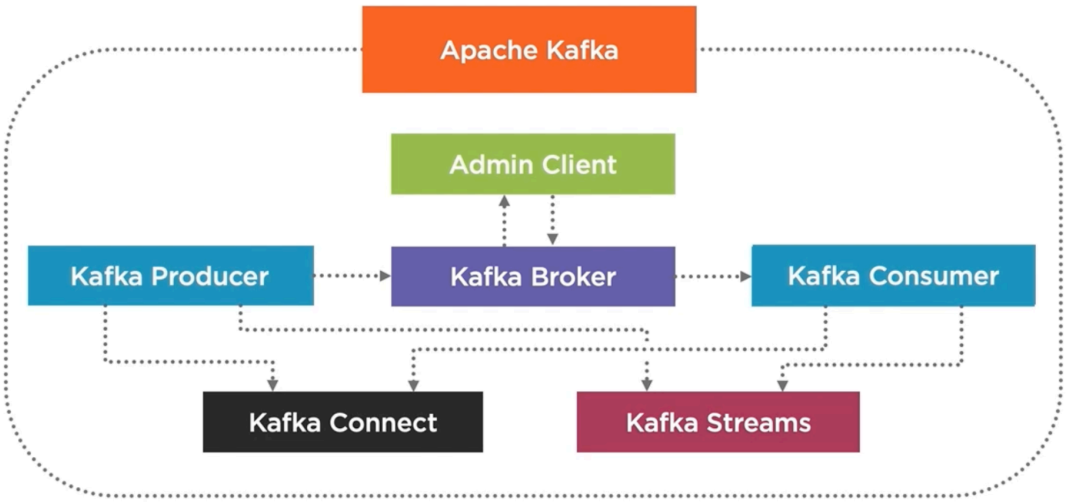
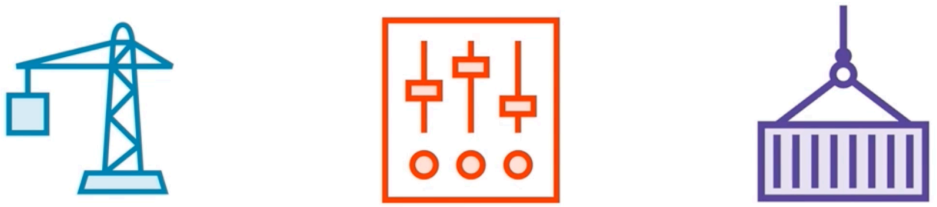
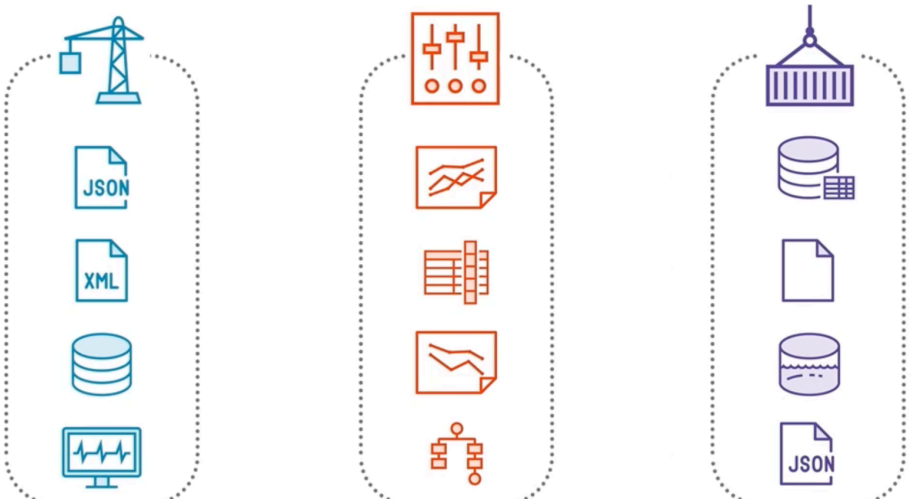

Big Data
A field that treats ways to deal with data sets that are **too large** or **complex** to be dealt with by traditional data-processing applications
- Wikipedia
Fast Data
A field that treats ways to deal with **streaming data** (data in motion) by allowing **instant** processing as it arrives in the system
- Me


Apache Kafka
Admin Client
Kafka Producer
Kafka Broker
Kafka Consumer
Kafka Connect
Kafka Streams

ETL



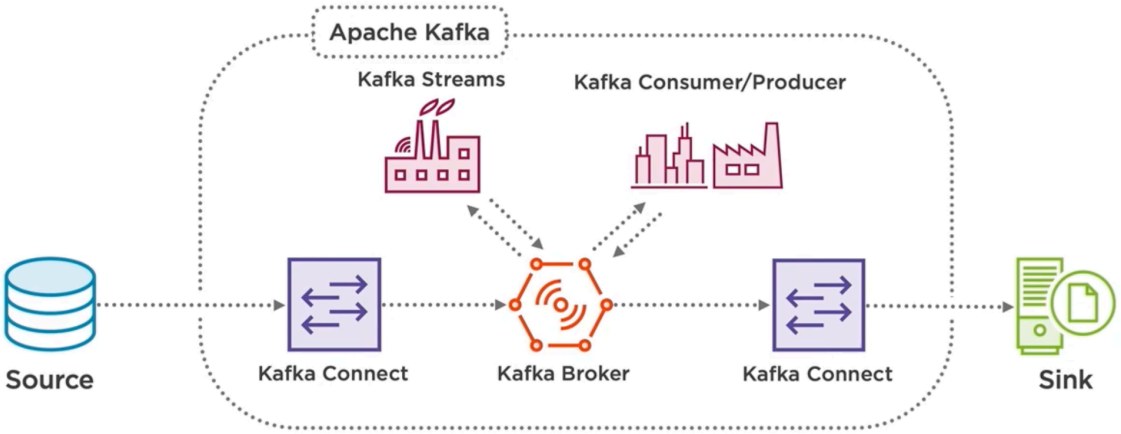
Extract Transform Load



ETL Tools

ETL Tools

ETL with Apache Kafka



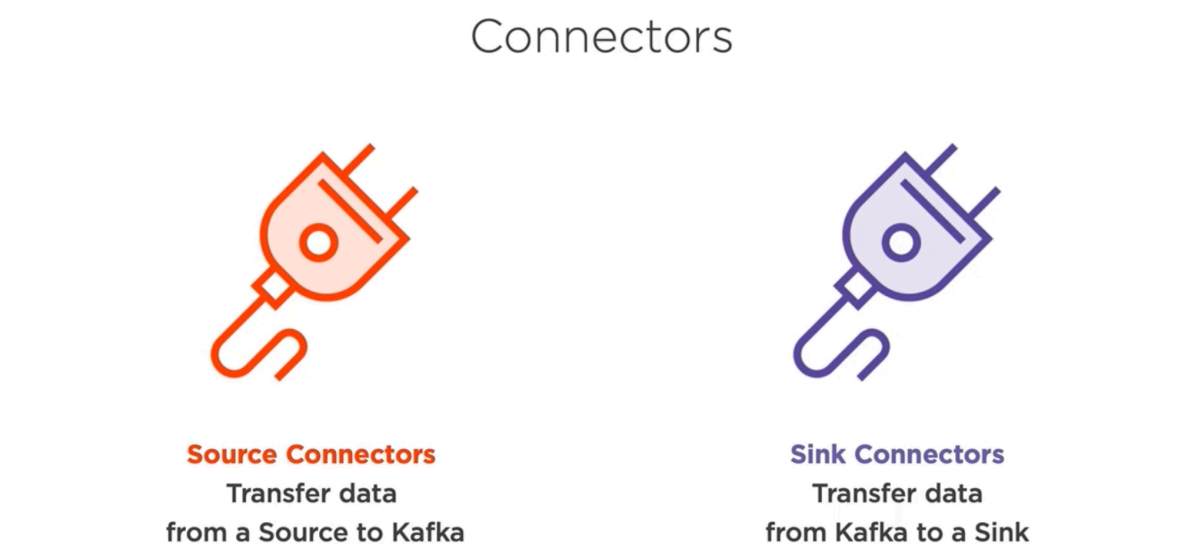
Kafka Connect

Is **NOT** an ETL tool by itself! It has some ETL capabilities but it needs to be integrated with some other components from the Apache Kafka Ecosystem.

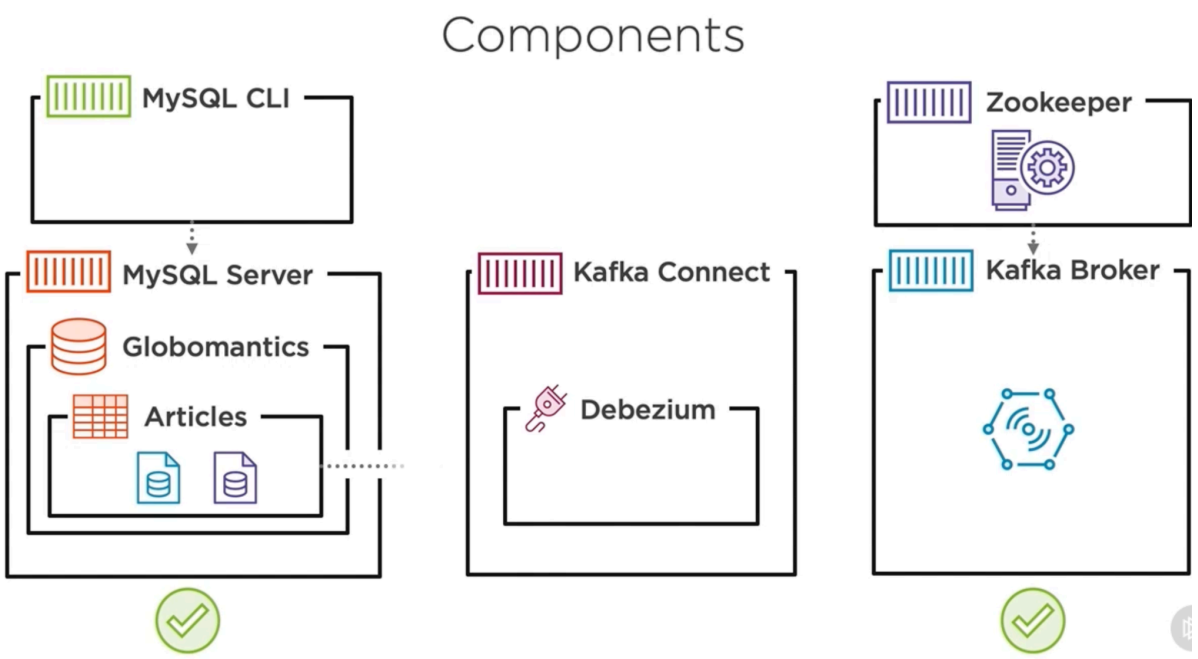
Apache Spark vs Kafka Connect

	Apache Spark	Kafka Connect
Extract	✓	✓
Transform	✓	✓ ✗
Load	✓	✓
Dependencies	✗	✓

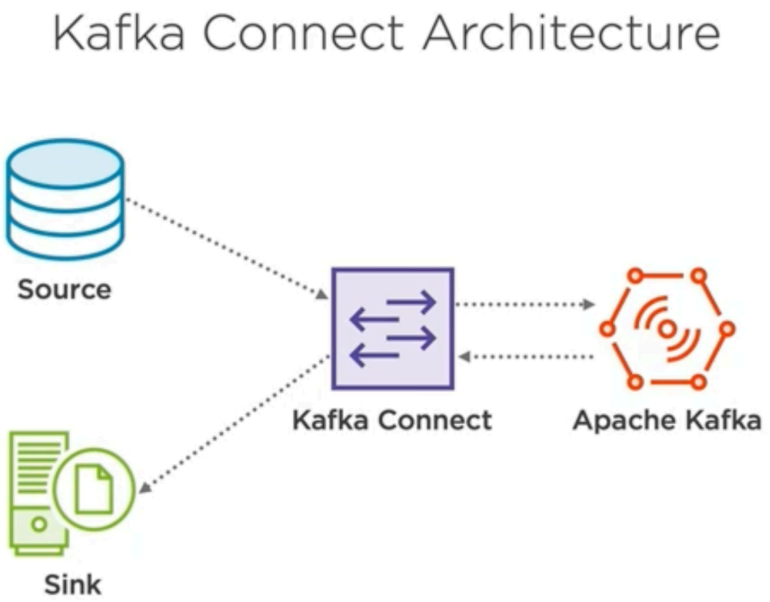
Data Flow


Connectors
Source Connectors
Transfer data
from a Source to Kafka
Sink Connectors
Transfer data
from Kafka to a Sink

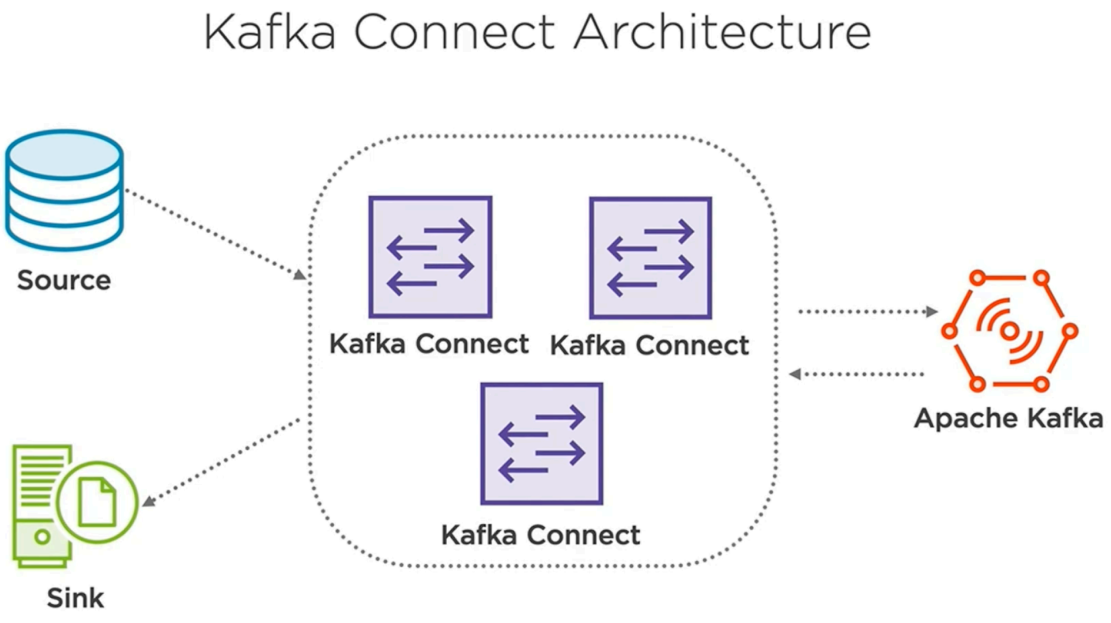
Demo:


Components
MySQL CLI
MySQL Server
Globomantics
Articles
Kafka Connect
Debezium
Zookeeper
Kafka Broker

Single Instance Kafka Connect



Scaling Kafka Connect:

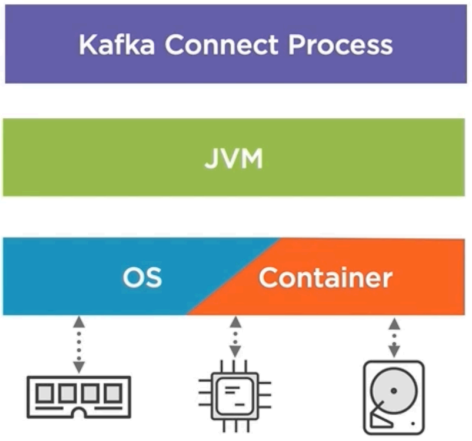


Kafka Connect

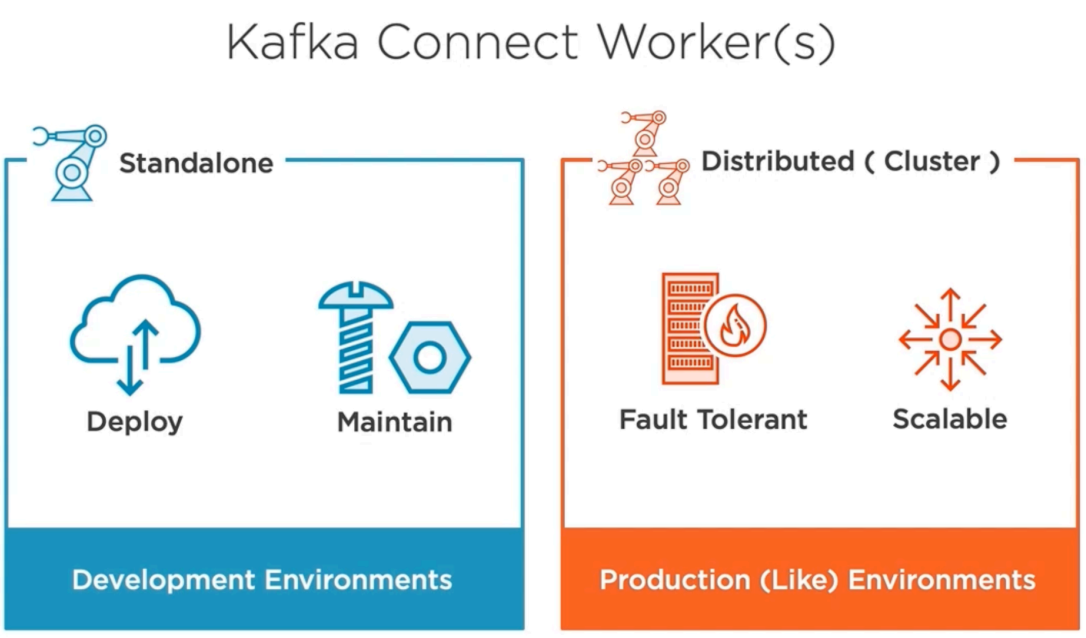
Streaming

Batching

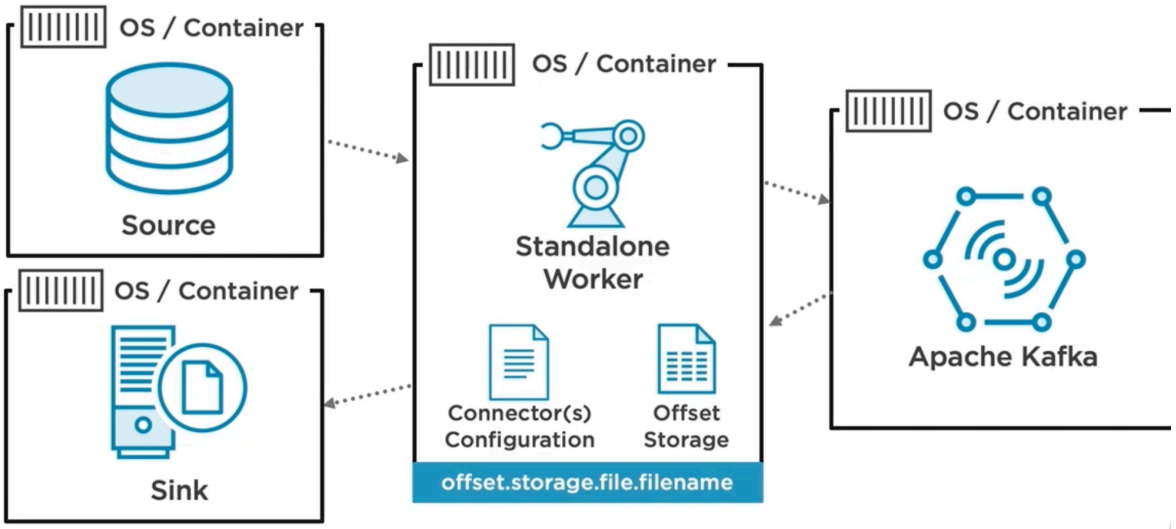
Kafka Connect Worker



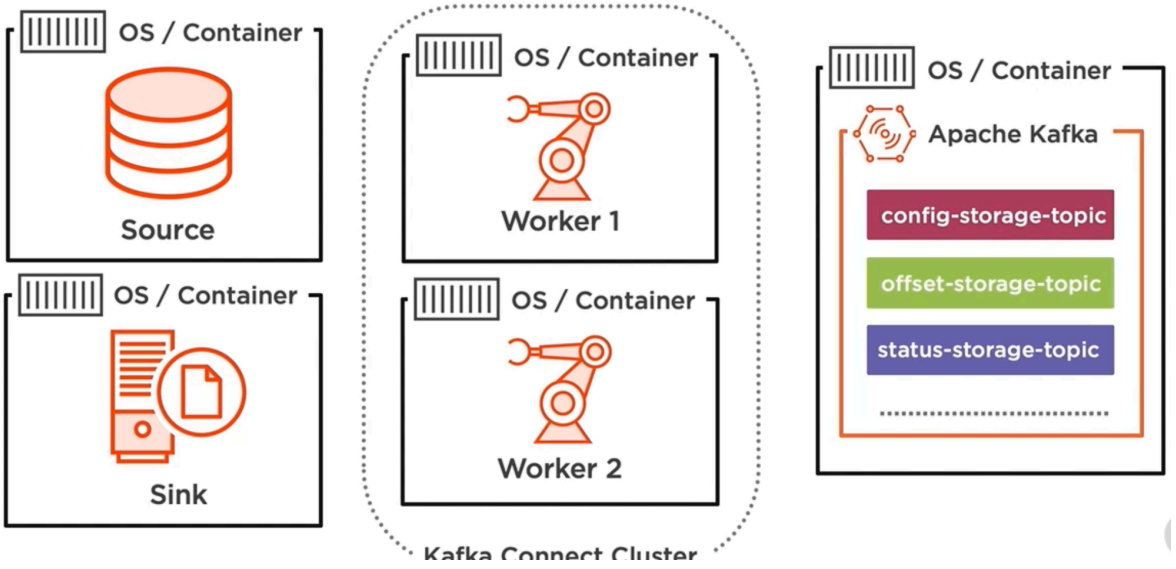

Standalone
All work is performed in a single process
Distributed
Work is shared and balanced across multiple nodes


Kafka Connect Worker(s)
Standalone
Deploy
Maintain
Development Environments
Distributed (Cluster)
Fault Tolerant
Scalable
Production (Like) Environments

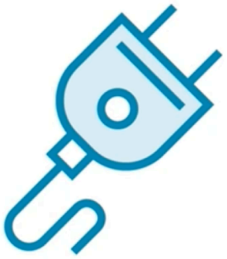
Kafka Connect Standalone



Kafka Connect Distributed

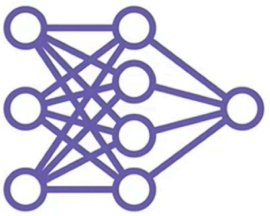


Connectors



Connector (Plugin)

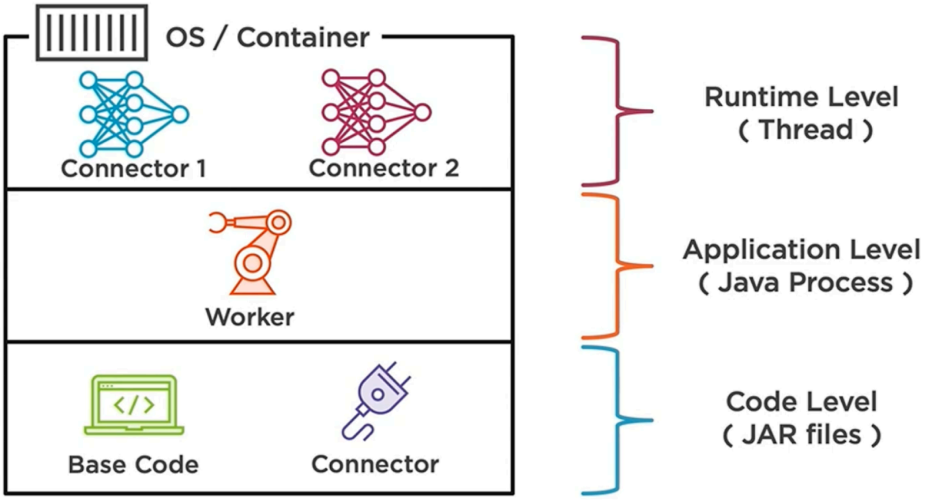
An **adapter** that allows connecting to/from a common system



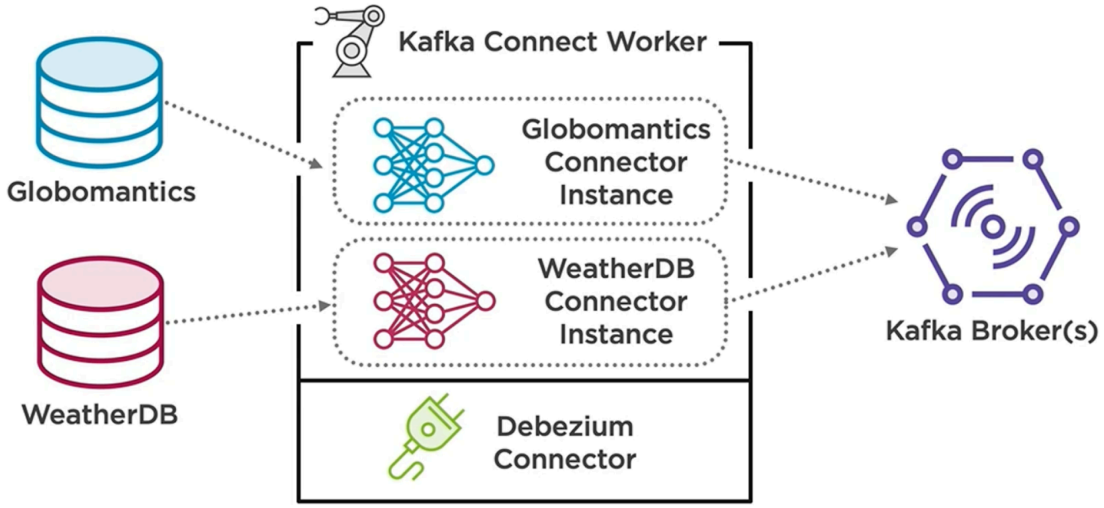
Connector (Instance)

A **job** that enables the exchange of data between Kafka and other systems

Connectors



Connectors



Connectors