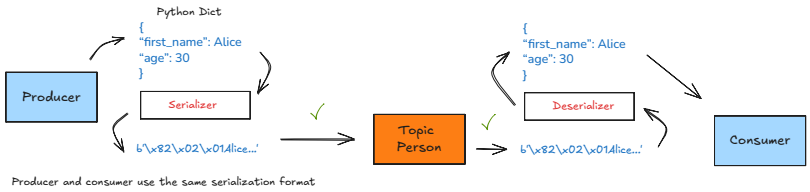
Serialization

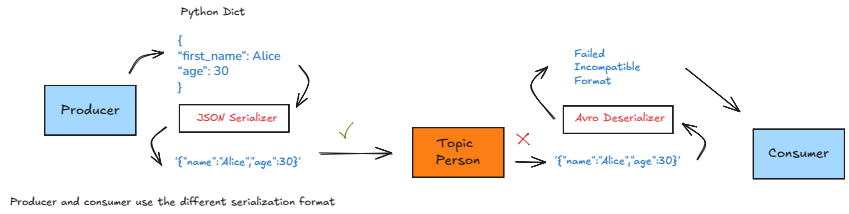
- Before fixing schema drift, let's talk about serialization



What is serialization?

The process of turning data structures into bytes so they can be stored or sent through the network.

Kafka stores only bytes
It doesn't understand Python objects



Why does it matter?

Serialization is the language that both sides agree on, the contract of how bytes map to fields.

Without a consistent rule, producer and consumer might interpret the same bytes differently

Serialization Formats

Common formats:

FEATURE	JSON	AVRO	PROTOBUF
Type	Text-based	Binary	Binary
Readability	Human-readable	Not human-readable	Not human-readable
Speed & Size	Slow, verbose	Fast, compact	Fast, compact
Schema	Optional / none enforced	Required and stored separately	Required and compiled into code
Schema Evolution	Difficult – no versioning	Excellent – built for evolving schemas	Limited – field numbers must remain fixed
Best For	Simple debugging, APIs	Streaming & data pipelines	Inter-service RPC (microservices)
Compression	Poor	Very good	Very good
Language support	Universal (text)	Broad (multi-language libraries)	Broad (multi-language code generation)
Typical use	Logging, configs, REST APIs	Kafka, data lakes, schema-registry pipelines	gRPC, internal APIs