

Apache Avro

Presented by Riaz Farhanian - August 2022

What is Apache Avro?

Avro

 Avro is an open source data serialization system that helps with data exchange between systems, programming languages, and processing frameworks. Avro helps define a binary format for your data, as well as map it to the programming language of your choice.

Avro

- Avro has a JSON like data model, but can be represented as either :
 - JSON
 - Compact binary form
- It comes with a very **sophisticated schema** description language that describes data.

Avro Schema

```
"type": "record",
"doc": "This event records the sale of a product",
"name": "ProductSaleEvent",
"fields" : [
  {"name":"time", "type":"long", "doc":"The time of the purchase"},
  {"name":"customer_id", "type":"long", "doc":"The customer"},
  {"name":"product_id", "type":"long", "doc":"The product"},
  {"name":"quantity", "type":"int"},
  {"name":"payment",
   "type":{"type":"enum",
           "name": "payment_types",
           "symbols":["cash","mastercard","visa"]},
   "doc": "The method of payment"}
```

Avro JSON

```
{
  "time": 1424849130111,
  "customer_id": 1234,
  "product_id": 5678,
  "quantity":3,
  "payment_type": "mastercard"
}
```

Why use Avro with Kafka?

- 1.It has a direct mapping to and from JSON
- 2.It has a very compact format. The bulk of JSON, repeating every field name with every single record, is what makes JSON inefficient for high-volume usage.
- 3.It is very fast.
- 4.It has great bindings for a wide variety of programming languages so you can generate Java objects that make working with event data easier, but it does not require code generation so tools can be written generically for any data stream.
- 5.It has a rich, extensible schema language defined in pure JSON
- 6.It has the best notion of compatibility for evolving your data over time.

Let's use Confluent!

- At the core of Confluent Platform is **Apache Kafka**, the most popular open source distributed streaming platform.
- Confluent Schema Registry provides a serving layer for your metadata.
 - It provides a RESTful interface for storing and retrieving your **Avro**®, JSON Schema, and Protobuf schemas.
 - It stores a **versioned history** of all schemas based on a specified subject name strategy, provides **multiple compatibility** settings and allows **evolution of schemas** according to the configured compatibility settings and expanded support for these schema types.
 - It provides serializers that plug into Apache Kafka® clients that handle schema storage and retrieval for Kafka messages that are sent in any of the supported formats.



