Efficient Serialization of Spatial Data

Geospatial Sensor Webs Conference 2017
Malith Withanage Don

What is GSoC ?

- Google Summer of Code (GSoC) is an online, international program organized by Google which happens annually every summer
- Encourage university student participation in open source software development
- Help open source organizations to identify and bring in new developers each year

- This year, Me + 

+ 

> Efficient Serialization of Spatial Data

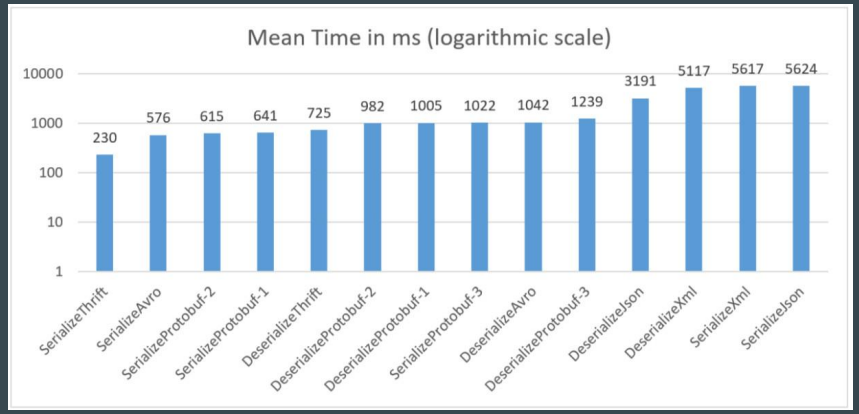
What motivated us ?

There is a lack of adaptation to modern serialization frameworks in Geo Spatial space.

Binary and Text serialization techniques locks you into one programming language and language-agnostic formats like XML, JSON are too verbose and too slow to parse.

There are common frameworks for binary serializations, but there is no common approach to use these for spatial data.

Reference - http://labs.criteo.com/2017/05/serialization/



XML / JSON

What is the benefit ?

Spatial Data is growing at a rapid pace and following operations needs to be performed to facilitate various business critical applications.

- **Collect data** - Low latency data transfer over the wire and reduce network bandwidth is important
- **Process / Exchange data** - Data streaming and real time processing is critical in modern day business
- **Store / Archive** data - Efficient Serialization / Deserialization of data will drastically reduce the resource and space usage

How to fulfill these requirements?

What is our solution ?

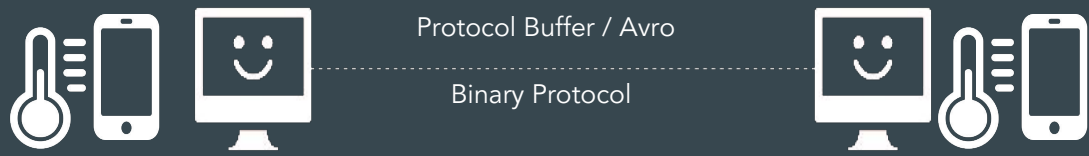
GitHub

Source code is licensed under
Apache Licence 2.0
Available in github ->
topology-serialization-framework

Using Protocol Buffer and Avro,

- API for (de-)serialization of vector and raster data.
  - Schemas to represent vector data by following the OGC Simple Feature Specification
  - Schemas to represent raster data by following the ISO Coverage Model.
- Benchmark the performance in accordance with time and storage space to identify comparison metrics.

Why we chose it ?

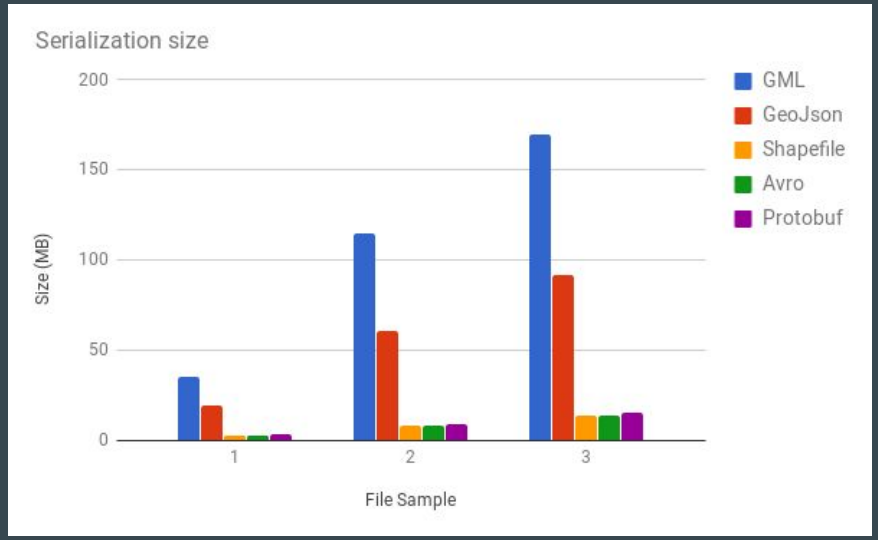


Protocol Buffer and Avro offer,

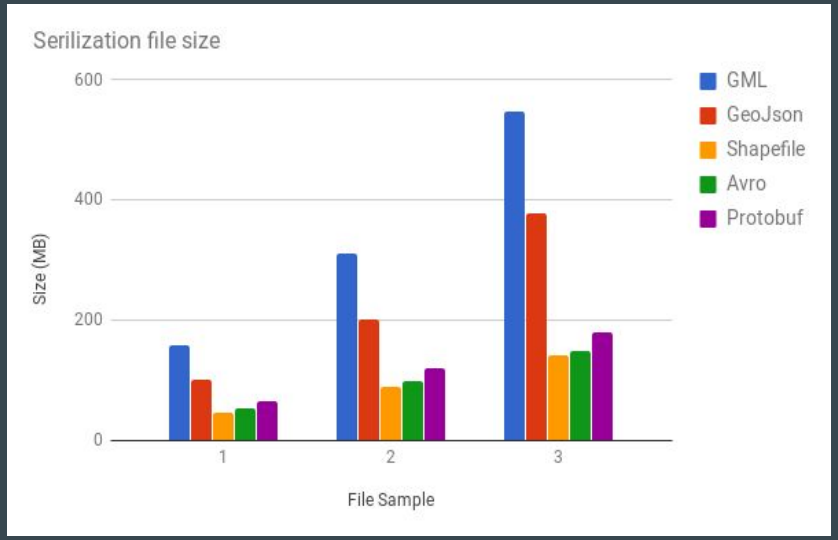
- Better performance in the context of CPU and memory needed for processing, limited network bandwidth and storage.
- Language and platform neutral.
- Supports schema evolution to deal with fluxing behaviour of data and has built in RPC support.
- Heavily used in, Apache Hadoop for data streaming and storing, Facebook and Cassandra as a low latency network communication protocol and in Google for efficient data encoding and decoding.

Benchmark results - space

Comparing file sizes of different techniques to represent same raster data set



Simple Structures (Points)

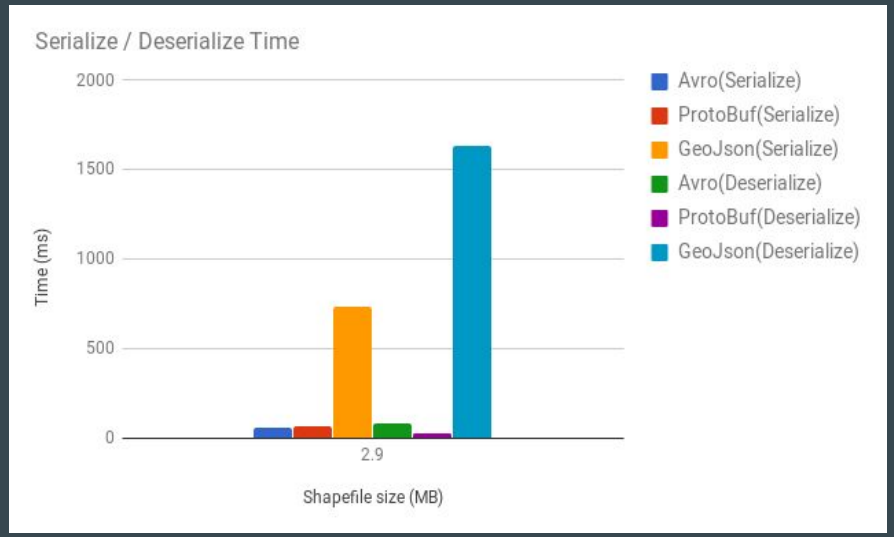


Complex Structures (MultiPolygons / MultiLineStrings)

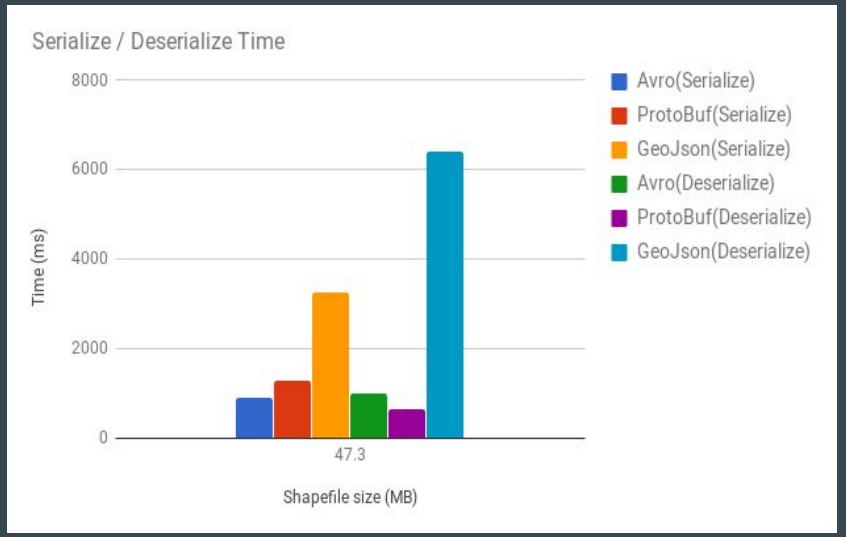
Reference data set - http://download.geofabrik.de/europe.html (Belgium, Netherlands, Spain)

Benchmark results - time

Comparing Serialize/Deserialize time of Protobuf and Avro vs different raster data sets



Simple Structures (Points)



Complex Structures (MultiPolygons)

Excluded GML -> Deserialize 26193 ms

Reference data set - http://download.geofabrik.de/europe.html (Belgium, Netherlands, Spain)

What we found ?

- Bandwidth and time for (de-)serialization will be drastically reduced when using Avro and Protobuf compared to GML and GeoJson.
- Space used by Avro is less, and useful in archiving spatial data.
- Average Serialize/Deserialize time of protobuf is less compared to Avro, so it's useful in data exchange between servers.
- Discussion: But NOT a complete replacement for JSON or XML, especially for services which are directly consumed by a web browser because these are in binary format.

Outlook

- Define schemas for other serialization frameworks such as MessagePack, Flat Buffers,Thrift etc ...
- Profile some other performance metrics such as network latency, memory and CPU usage etc ...
- Utilize these schemes and implement serialization API to serialize Geo Spatial Data to prefered destination formats (ie - JTS or Coverage Model)
- Integration of serialization API to real world applications such as the Sensor Observation Service or the Web Processing Service for efficient serialization of Spatial Data