# Summary

Data Serialization – Avro & Sequence Files is the project which I chose.

The main objective of this project is to learn about the two most important serialization mechanisms provided in Hadoop and work on some basic examples.

Learnt about Avro and Sequence Files how it is implemented, importance of the two key parameters i.e. time and space in the distributed environment.

Hadoop‘s native library provides Writables for data serialization (converting object data into byte stream) and deserialization (converting byte stream data to object data) and also it provides support for Sequence Files which will store the data in binary format. These are the only two mechanisms provided by hadoop for data serialization.

The main drawback of these two mechanisms is that Writables and SequenceFiles have only a Java API and they cannot be written or read in any other language. So any of the files created in hadoop with above two mechanisms can’t be read by any other third language which makes hadoop as a limited box. To address this drawback, Doug cutting created Avro, which is a language independent data structure.

AVRO is a data format created by Apache to support the Hadoop eco system. Other data formats are XML, JSON, Protocol Buffers (by Google), and Thrift etc.

Avro stores data by putting data definition with the data allowing for Avro files to be read and interpreted by many different programs. It stores all of the data in a binary format making the files more compact, and will even add in markers to help Map Reduce jobs find where to break large files for more efficient processing.

The challenges I faced was the time as I work full time, getting a dedicated focus after work has been challenging in this course.