Gurjus Singh

MSDS 432 Foundations of Data Engineering

July 19th, 2020

Module 4 – Reading Comprehension

1. Discuss some of the problems with JSON, XML, and CSV file types. (2 pt)

Common problems between JSON, XML and CSV formats is firstly, XML and CSV formats cannot distinguish between strings and numbers, and JSON cannot determine the difference between integers and floats and cannot specify precision. Secondly, JSON and XML do not support binary strings. Binary strings are different from text because they are data that support pictures and consists of 1s and 0s. Thirdly applications that do not support XML and JSON schema have to be coded in that schema. A schema is defined as a structure. In terms of CSV it does not have a schema, and there is no clear definition on what the rows and columns can contain.

2. What is Thrift and Protocol Buffer, and how is it different than Avro? (4 pts)

Thrift and Protocol Buffers are two encoding packages developed by Google and Facebook. Encoding means to put data in a specific format. They both require a schema. They are both similar to each other such as having a body where the data is contained surrounded by curly brackets. The key difference is when defining in Thrift a struct keyword is used. Syntax includes colons and commas, while in Protocol Buffers equals and semicolon syntax is used. Another key difference is Thrift as two formats which are Binary Protocol which fits data in 59 bytes and Compact Protocol which fits the data into 34 bytes. Protocol Buffers differ because they only have one format to fit the data in and can fit it in 33 Bytes. Avro is different from both Thrift and Protocol Buffers because firstly it is the most compact with only 32 bytes to fit the data. Avro also has concepts known as the writer's and reader's schema. What this mean is that the application can write data in whatever structure it wants as long as it is compatible and can read data in whatever structure it wants.

3. In regards to RPC, how is a network request different than a local function call (2 pts)

Key differences between network request and local function call are: a local function call can either succeed and fail which means it is predictable and under the user's control, while a network request can be unpredictable due to out of control causes such as the machine is unavailable or the request may be incomplete. Also, for network requests, sometimes the request might have been successful, but the response from the machine receiving the request might have gotten lost. This request might be duplicated if this happens. A local function call speed is faster than a network request which has latency issues. With local functions you can pass data to pointers to store in local memory while with network requests the data must be encoded, but this can be a problem if the data is large.

4. What are message brokers? Give some examples. (2 pts)

Message brokers help send a request or message to another process. For example, it can help pass the message at the right time to a process on a different machine. It can also resend the message, just in case it is lost. It can also send the message to several destinations. It does not need the user to know much about the destination. Some examples of message brokers are RabbitMQ, ActiveMQ, and Apache Kafta. The way Apache Kafta one of the message broker examples works is that it stores messages in key-value pairs. These messages come from other machines called “producers”.