# Requirement

Create a system that can read and process 3 log level data sets (CSV formatted) in a way that you can do the following:

* Join the data sets across all 3 log files to create a pseudo “log”
* Filter/Sort based upon particular columns (eg filter where HTTP Status Code is 200) and getting a count of the rows
* Support taking a sum or avg of numeric columns (eg give me the avg Response Time (MS) for all routes)

#### CSV Data Sets

Entity DB dump Log Contains a set of an object/entity table

Error Log Contains a set of errors that have occurred within our system

API Endpoint Log Contains a set of data explaining metrics around the api calls to a given set of endpoints within our system

**Task is to solve following business initiatives based on the requirement**

1. Which endpoint has the most internal server errors (eg. HTTP Status Code = 500)

2. When endpoint generates the most errors in the error log?

3. What is the fastest (avg response time in ms) API endpoint?

4. What is the slowest (avg response time in API endpoint?

5. What the average response time for the “/entity/find/” endpoint?

6. What Entities have nothing in stock?

7. What is the average amt for all Food types AKA entity parent\_id = 1

8. Which Entity gets accessed the most via the API?

9. What day had the most api traffic?

10. What day had the most errors?

# Design

Considering the structured log files, used Spark SQL that is a Spark module for structured data processing. Used Datasets and DataFrames concept and developed code in Scala to solve 10 business initiatives.