# http://www.ridgebizdev.com/img/Infare_new_logo_black.pngData Platform Engineer test exercise

## Task 1: ETL/copy of data from S3 to DataBase of your choice

I’ve setup following environment for those tasks:

* ElasticSearch 6.2.2 with Kibana & ElasticSearch-Sql  
  I didn’t used ES before rather Teradata, GreenPlum or pure Hive but it is being evaluated at my current company so I choose it for learning opportunity
* Scala 2.11 with Spark 2.2.1 project built via sbt & and run locally in Intellij. I didn’t used Scala much before (rather Pyspark (Python 2.6.6) with Pycharm & Spark 1.6.2) but I choose it because:
  + Again wanted to use it a learning opportunity
  + There was already some Scala code provided
  + Based on some initial research Scala provides functionality required for the exercises (like easy integration with ES, error handling with Try & Either)

Code for the task is available in GitHub:

<https://github.com/MaciejGajewski/ETL_TestExercise/blob/master/src/main/scala/task1/SimpleDataLoader.scala>

It contains code to read csv data into rdd, wrap it in DataModel case objects, calculating additional week column, measure processing time and save it in a log file. Then it creates 2 types of indexes in EL:

* One with all the data in 1 index TODO: Provide metrics
* Second is a group of indexes partitioned by week TODO: Provide metrics

Due to limit of my laptop, access key to the files not working and time restrictions I processed only single file and couple of weeks.

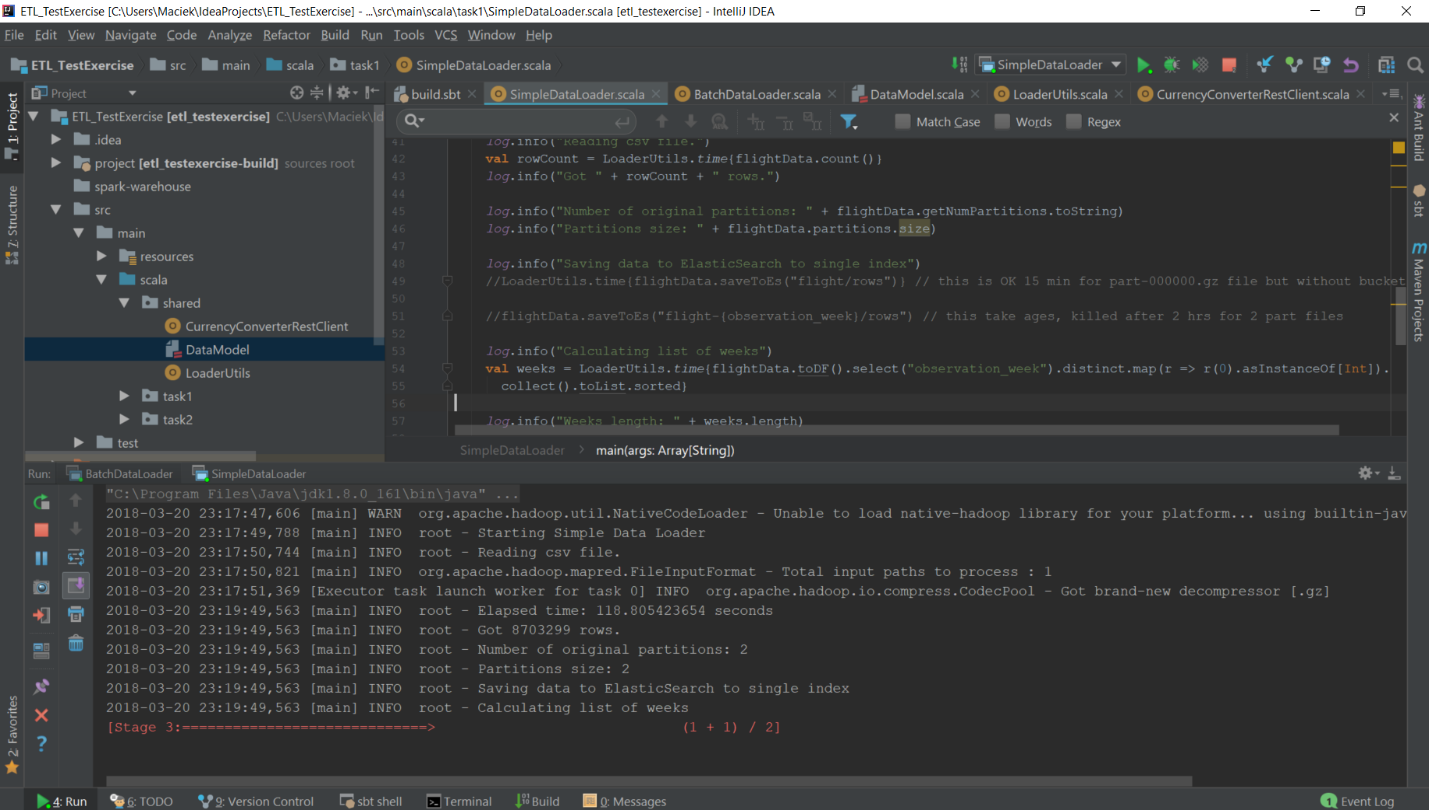
Having more time:

* I would try to run it on local Spark cluster via spark-submit or researched some free cloud environments.
* Enhanced code to calculate time metrics automatically

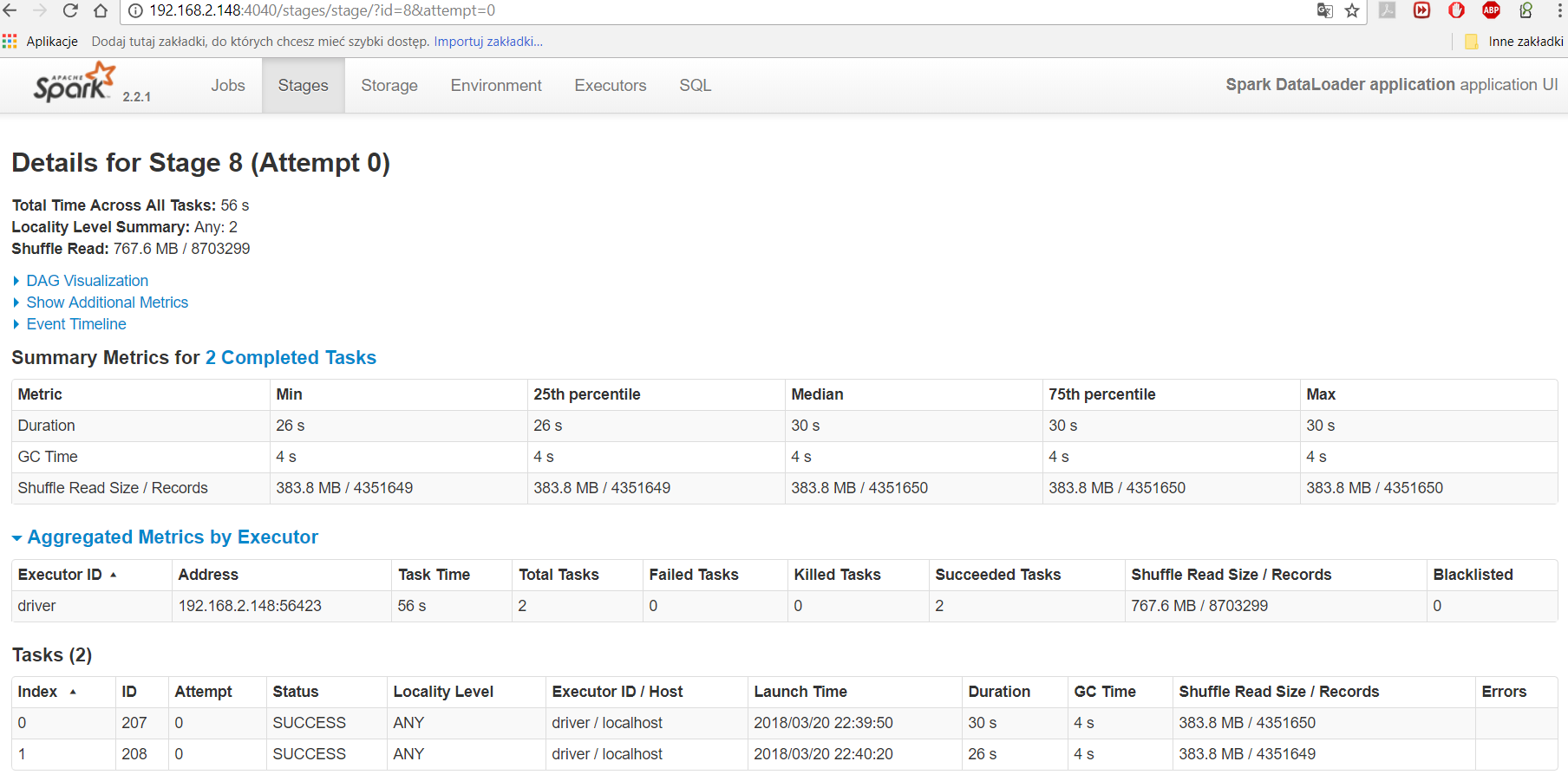
Other files are code for

* some utilities (data model, time metrics & web service call)
* logging config
* build config
* some excel calculation for a observed week based on existing number of days since 01.01.200

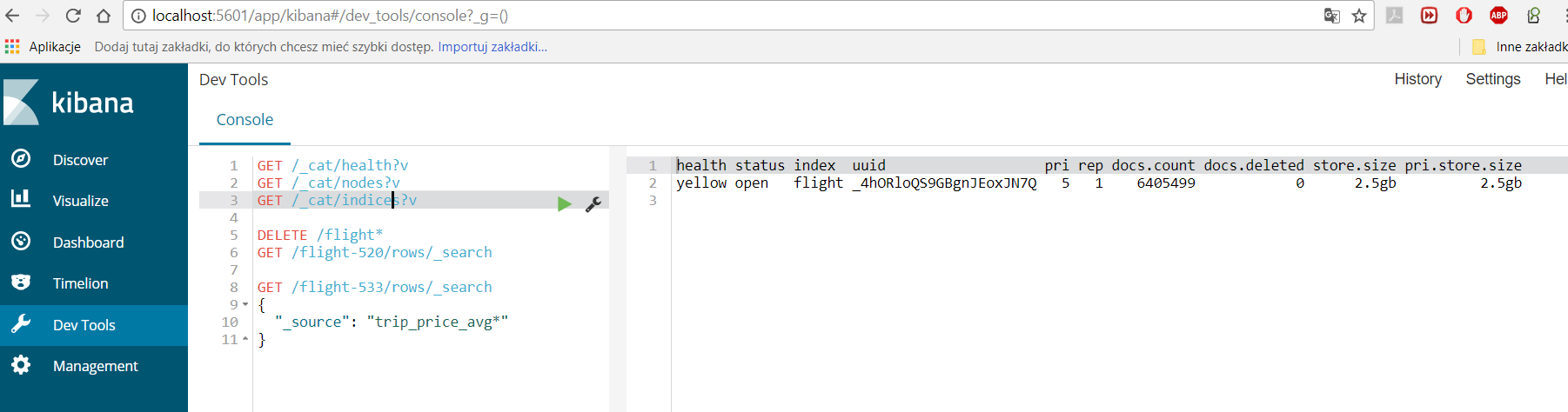
Intellij project:



Spark Console:



ES & Kibana:



## Task 2: Consuming Rest service during batched ETL process

Code for this task is available in <https://github.com/MaciejGajewski/ETL_TestExercise/blob/master/src/main/scala/task2/BatchDataLoader.scala>

It again reads csv file and wraps in DataModel cases objects. Then it prepares list of weeks but due to resource constraints I ran calculation only for couple of weeks. I don’t have much experience in building web services in Scala so I chose simple apache http client to make API calls to currency service.

Calls to current service REST api are done inside mapPartitions and map methods so they would be executed parallel if ran in multimode/multicore clusters. In order to manage service unavailability calls are wrapped in Try/Either object and successful attempts can be later filtered out and processed separately from failures.

If I had more time I would research proper/most robust way of building web service clients, added retries, waits etc.

## Task 3: (optional) Rest service exercise

Skipped this lacking experience in building web services in Scala or Python (I only once built web service client in Java using Spring).

## Task 4: (optional) Data mining exercise

TODO

## Task 5: (optional) Visualization exercise

TODO