UCLH PEACH and NHS Open Source: OpenEHR Architecture and Analytics

Team 38: Report 8

Sandipan Ganguly
Mengyang Wu
Desislava Koleva

Overview of progress :
Since the last couple of weeks was reading week and scenario week, we did not make any significant progress with our work. However, during the first week we continued working with our workflow on the local machines and implemented Apache Spark within the system. Therefore, as it stands now, the data flows in through NiFi where it converts the .csv files to small JSON files which is fed into Apache Kafka. Apache Kafka then sends the JSON files to Spark which queues jobs for a batch of files.

We also set up a couple of DC/OS servers to test and implement our local workflow onto Azure.

Problems Faced:
After the deployment of a new DC/OS for testing the completed workflow, we started to repeat the task we did on our local machines. However, problems occurred in all the components. Nifi can be installed successfully but we cannot access the design UI as we did locally before; The installation of Kafka cannot be finished with unknown errors; Although Spark can be installed smoothly, it cannot come into play without the preprocessor and the message hub.

Successes:
The only success throughout these two weeks was finishing the completed workflow (Nifi - Kafka - Spark) locally, which is the basis for us to repeat the deployment process on DC/OS.

Plan for next two weeks:

No.	Task
1	Implement our workflow onto the live DC/OS Server on Microsoft Azure

2	Install Druid and start preparing to implement that into our local workflow
3	Look into Kafka Streams and Elastic Stack to add on to our workflow
4	Look towards improving the Kafka streams processing speed

Summary of meetings held:

Meeting Date	Who attended	What we did
13/2/17	Sandipan Mengyang Desislava	● Went over the previous team's Apache Spark installation and discussed how we can implement that in our system ● Created a couple of DC/OS servers to attempt and deploy our workflow on Azure
17/2/17	Sandipan Mengyang Desislava	● Discussed the roadmap for the next couple of weeks and what we would prioritize after Apache Spark ● We decided on Druid and Kafka streams since they were important in the previous team's workflow

Individual Contributions:

Sandipan Ganguly

During the first week, I looked into the previous team's Spark installation and did some research of own on how it can queue JSON files from Kafka. I also attempted to install the Confluent Kafka stream connectors to our workflow but was unable to do so due to the lack of documentation from the previous team. . Furthermore, I also looked into ways to implementing the entire workflow into DC/OS starting with NiFi.

I wasn't able to make much progress other than going over the future roadmaps, due to the following week being Scenario Week.

Mengyang Wu

During the last two weeks, my main contribution is completing the workflow locally and attempting to deploy DC/OS with all the components running. The unfinished step two weeks ago was building the pipe from Kafka to Spark so I focused on implementing that process. The outcome is good and we finally have all the methods to generate data, preprocess data, send data and use the data for future machine learning jobs. However, the next step to make the remote deployment is an adventure with a lot of difficulties, even with well-defined documents of DC/OS, some issues cannot be addressed because our special running environment (Azure). Therefore, in the following week, I will continue working on the deployment. Hopefully we can overcome all the problems.

Desislava Koleva

Over the last two weeks, I was unfortunately unable to make much progress on the project, as I was primarily focused on work related to other university modules during reading week, and on our extensive scenario week assignment during scenario week. This week, I have started briefly looking into Kafka Streams and how to implement it on my local machine, as well as continued familiarising myself more with the Microsoft Azure DC/OS environment in general. My aim is to make significant progress on these tasks within the next few days so as to be able to help implement the entire workflow onto the DC/OS server in the upcoming weeks.