Project CS 2017 Course Feedback

Uppsala University

Fatimah Ilona Asa Sabsono
Daniel Edin
Filippos Petros Lanaras
Emanuel Lind
Martin Matus Saavedra
Michael Wijaya Saputra
Rahul Setty
Satya Vrat Shukla
Ludvig Strömberg

January 12, 2018

Contents

1 Project Management 3

1.1 Scrum 3

2 Feedback for Scila 4

3 Feedback for the University 4

Appendices 6

1 Project Management

1.1 Scrum

Scrum is an agile methodology for software development and we decided to use it for this project. Scrum is dividing the project into several Sprints, a length of time that would consist of the whole software development life cycle with certain predefined goal, to make it more manageable in tracking the project progress and resolve problem faster. For each sprint we have sprint planning, daily scrum, sprint review, and sprint retrospective.

Sprint planning is usually done at the first day of a sprint. We define the goals or Minimum Viable Product (MVP) for that particular sprint, breakdown the task, who will gonna do those task or be responsible for it, how long each task will take time, and also improve our performance based on previous sprint. In order to keep track of progress, we have daily scrum which is a meeting taken place before everyone working to talk about what problem each person have, what they did yesterday, and what will they do today. We also have sprint review at the end of each sprint where we present the MVP to Scila AB for that sprint and get feedback from them. After that, we do a sprint retrospective where we evaluate our performance internally among team members based on what we achieved, Scila AB feedback, and how much the workload for us.

We use an online tool and offline tool to do Scrum. We use Trello as an online collaborative tools to store and keep track of everything that has been done and to inform project course coordinator and teaching assistants. Trello is so easy to used if we are not in the room, but it is not very suitable for us because we always did the project in the class. Thus, Trello only serve as way of communicating to people outside of team and storing task. We rather prefer to use sticky notes and a white board to keep track of all the task for the current sprint. It is more up-to-date and easier for us to use.

2 Feedback for Scila

First of all, we would like to thank you for this chance of working with Scila AB and getting to know current technology in financial technology industry through them. This project is a good idea of introducing what kind of real problem that they are facing, what is the current solution for that problem, what could be explored in the future, how stock market works and technology related in it. Although it has been such a good experience and we really grateful for the fruit basket, we would like to give our feedback for Scila AB.

- The specification stated that we would work with big data but event obtaining some data too much time and it was random and small size. It required a lot of time from us to duplicate the data and of bad quality.
- It was great to get input from anyone at Scila, but maybe stick to the big picture of the project would be great instead of giving new idea in each meeting/feedback.
- The person in charge need to know exactly what kind of objective should be achieved, in that way they could help us to stay in track.

3 Feedback for the University

We would like to thank you everyone from the university to give us this experience through Project CS course. This project give us a chance to work with Uppsala University cloud and have a good knowledge about Scrum. Our feedback here would be about all the the things that could have been better in running and supporting this course. The teaching assistants could be more helpful if they actually understand the specification or input from Scila.

- The interface of the OpenStack can be improved to be able to create a new flavor without the need of contact the stuff to do that for us. This would help us experiment even more with the specifications of the VMs.
- The current OpenStack infrastructure takes forever to spawn new VMs from our snapshots which slowed us down while creating them.
- The networking of the VMs is quite strict and inconvenient to work with. It would have been helpful if there a guide or a way to connect our computers to

that network in a fast and convenient way.

- The course administration should have kept a steady front to the attendance in the course.
- Some of the hardware provided by the IT department were not working very well/ half-broken, i.e. a few keyboards and mice.
- There should have been proper high back chairs for all of us and without specific requests. This caused some of the team members to have back pains.

We would like to thank Edith, Amendra and Xiaming for the input and feedback. We would like to thank Edith for providing us with free coffee and fika and sponsoring our social activity (laser tag) in the beginning of the course. Amendra's critics in the machine learning modules was very helpful, regardless that initially felt like being blunt it, it was a great input and helped us to get better.

Appendices

A: Contributions

Fatimah Ilona Asa Sabsono

My contribution in general for this project is becoming the scrum master and the person that facilitate communication between the team, Scila AB, and University (project course coordinator and teaching assistants). As scrum master, I was in charge of making sure that every day at 9 in the morning we had a daily scrum and facilitate all scrum planning, scrum reviews, and scrum retrospectives. I manage all the meeting with Scila AB and University, make sure that all the predefined schedule is fulfilled, communicate the problem and resolve it with both sides, and make sure feedback is received by all members in the team.

My early contribution in this project was setting up the ground rule for the project management, give input based on my previous experience and available knowledge, and setting up the communication and management tools for the team. I mostly worked on the machine learning component for the rest of the time in this project by making the general data transformer in Spark, doing the whole classification approach in Spark, and coordinating with other members about machine learning approaches. For the machine learning part, I work with Michael, Rahul, Emanuel, and Ludde to discuss the implementation and I work with Satya and Daniel to discuss the available parsed data. I implement a visualization using an available Java library but it has too many limitations to be used for analyzing the data and we remove it from the application in the end. I also implement a module for generating a report in CSV format to be used in Tableau. I work with Filippos and Michael when designing the cluster. I am one of the people that were presenting in the mid-term presentation and tracking all the reports along with Satya.

I learn a lot during this project, especially about teamwork, project management, and how to be a person in charge of connecting everyone. I also learn more about technical skills from others, how to do pair-programming in term of keeping the code clean and in line with specific code conventions, and how to manage a good working version of a source code. Although there are so many good things that I learned, there are a few thing that could have been better. We should have had a bigger data size and the specification document and suggestion should have been in aligned.

Daniel Edin

Like most of us did in the beginning of this project was to research a lot. I had no prior knowledge to neither machine learning, cloud computing or Apache Spark and very little financial experience. During the first weeks I mostly researched the features of Apache Spark to better understand how the API were to be used in a good way.

From there I worked together with both Philip and Rahul to see how we could set up a good foundation for the parsing to scale as good as possible. We decided to have each Scila Message Type as their own dataset and with that came a lot of code to properly map the specification into Java and Apache Spark. Me and Rahul decided to implement the parsing with two different approaches. The idea was to gain knowledge of them both in how they scale, what problems and benefits they come with. It turned out that the approach I started out with was both faster but also more modular. Which hopefully would mean it would scale better in the end. Together with input from Philip I worked a lot on the implementation to structure the parsing so it would be as easy as possible to extend. With this came the massive work to define and implement all different message types into this parsing structure.

The next step for me was to more thoroughly identify possible optimizations and how to leverage Apache Spark's full potential. I implemented both how we cache the data to the memory but also how we could diminish the problems we had with the JSON file structure. That pointed us to a different file format called Parquet. And thus the implementation of that began.

The parsing evolved quite a bit throughout the project with many configurable additions. I worked on how we could read the data in a more efficient way for both the machine learning work and the spoofing algorithm. We found out that they benefited from having the data in two different ways.

I also helped out in defining the spoofing algorithm we ended up using. There was a lot of discussion and problem solving trying to figure out how we could do it good but not too complicated. I helped out with writing the code for one of the filters that was the most trouble with. But also with my additional input we came to understand many different corner cases in how spoofing can occur and how to configure the algorithm appropriately. It helped a lot for us to be many involved in the spoofing algorithm because it turned out to be very complicated to properly define. Like the parsing implementation the spoofing algorithm evolved and changed many times during the whole project.

I have learned very much thanks to this project. Like I mentioned in the beginning I had very little knowledge in most of the areas we were to dive in to. I found out properly coordinated Scrum developing can actually work really good in groups of this size. But also how important it is to communicate with all team members. Overall I am very happy with how we handled most problems and how our problem solving skills evolved. It was a very good experience to have a project of this size that really lets us students to get a feeling on how the real world works.

Filippos (Philip) Petros Lanaras

My contributions in the project were various. I managed git related things, the VMs and their infrastructure, helped with the sprint planning and other bits and bobs that had to be done.

More specifically, I started reading more about git and then I setup our git repository on Github and managed that it throughout the course. This entailed setting up the workflow, reviewing pull requests and giving back feedback for clearer and more understandable code. Another thing was to help the colleagues to go through the merge and rebase conflicts while having interactions with all parts of the software to have a higher level understanding of the code, how it looks and works. I also helped Asa with some of the scrum duties, managing the sprint planning, guiding through what should be done generally in the course of the project and specifically on each sprint. I was one of the people that presented in the mid-term presentation at Scila AB about our cluster environment, Docker, Spring and demoed the application. I was a presenter on the final presentation as well, I presented about our cluster environment, system overview, Docker and Spring.

In the beginning I integrated the In the beginning I integrated the code convention with a tool that would check about that automatically and then integrated a Continues Integration (CI) tool the source code each time come one would push changes to Github.

I configured the Horizon environment, setup the VMs with the necessary software, was in contacted the IT support and retrieved the data from Scila. I collaborated with Michael to install HDFS and Spark and create a prototype image to deploy it to many VMs. There was also a lot of debugging and tuning of the applications to solve various problems. I was responsible to have the cluster up and running smoothly. I developed some scripts and configured some little automations to help us work faster and seamlessly between the different environments, i.e. development, cloud, docker through Apache Maven and Spring and managed some of the Spring beans with Asa. I had to do a lot of research about Spark, HDFS and Spring.

I read about docker what it is and how it works. Then I started creating some docker images to ease the setup process of the cluster within Docker. I researched about available visualization programs (Tableau and QlikView), what they do and how they work. I did some software architecture along side with Daniel and Asa, trying to refactor the code while the software expanded. Apart of that I collaborated with Daniel with the parsing of the data, with Ludvig about DL4J, Asa with restructuring the machine learning part of the application and with Martin with the Spoofing. Coding various library classes and creating tests for them.

Through out the course I my improved communication, presentation and planning skills. Extended my scrum, working in a team and pair programming, finance, git, Spark, Spring, machine learning/models skills.

I liked that we had to work with a company with a real problem and had the opportunity to go into finance markets but at the same time I could feel this was still a course.

Emanuel Lind

At the beginning of the project I spend most of the time to do research. I read about weak and strong scaling, the differences about RDD, Dataframes and Datasets in spark and about the machine learning in spark. And I also set up my computer with the software needed and read about all the tools and framework I needed to know about to be able to work with this project.

I set up the Google Calendar for our meetings, presentations, when people were away from the lab and other things that we needed to put in a calendar and a Google Drive for our documentation. I also created a document for code convention and looked at how the testing of our code would be done.

After this I was mostly working with the machine learning part where I did the unsupervised learning together with Michael. We used clustering algorithms and that was used to address the third part of the project specification about anomaly detection. The work with clustering included making a model and test it with different input features, finding out the correct number of clusters for this data and finding ways to determine if a data point is an anomaly or not. I implemented the K-means and Bisecting K-means algorithms but decided that we would only use K-means because Bisecting K-means was slower and the result was similar so it was not worth the extra time it took to run. We had a lot of problems with this part because of the data that we had. It was not really suitable to do this kind of clustering analysis. So our final implementations focus mostly on statistics of cluster instead of finding anomalies.

Other Machine learning parts that I helped with was the preprocessing and visualization. I also started implementing feature selection using both what existed is Spark and looking at other approaches. It was later decided that we would choose our features by ourselves instead so that part was dropped.

By taking this course I learned a lot about Scrum, Git and got more experience in working with a team. I got more skills in machine learning and learned about new tools and frameworks like Spark, Spring and OpenStack. I also learned some things about the financial market and spoofing. At the mid term presentation I was one of the people who presented and I talked about the clustering part. So I improved my presentation skills as well.

Martin Matus Saavedra

The first two weeks of the project were used to research about spark, how it scales and how it should be used for optimal performance. The research that was documented and shared with the whole group. It was necessary and gave a broad view of how spark and all the functionality works. This also gave us the time to set up our computers with operating systems and software that was needed.

I then did most of my work within the spoofing group, creating a solution for component two. The parsing component is one of the main components that spoofing needs. Me and Ludvig began writing generic SQL queries to start off component two while also supporting the parsing group.

During the next few sprints, me and Satya started constructing algorithms for the different filters and parameters required for the spoofing detection. We followed the suggested parameters mentioned in the project specification provided by Scila AB.

During these sprints we had many questions regarding the implementation of the spoofing algorithms as we did not understand some of the definitions. We had multiple meetings with Scila to clarify and get feedback. The implementation was changed accordingly. After this I worked with Daniel and many improvements were done. The performance and algorithms evolved with time.

Results were written to csv at first to ease the work of manually checking orders and trades. It was however switched to write datasets to json to match the scila alert message and running the data day by day instead of the whole data folder at once. This was done and to make the deliverables to Scila.

Michael Wijaya Saputra

In the beginning of the project, I have done research and study literature about spark, HDFS, and financial things. However, I have done more research and study literature about HDFS and Spark. In this time, I have learned about the scala language. Later on, I and Philip have worked together to prepare a cluster in our cloud environment. Most of time, I have worked to prepare HDFS to run in our cloud environment. For a while, I have done maintenance for HDFS and implement spark history server in our cloud environment and make a documentation and maintenance file for HDFS.

Afterward, I have made code which could be used for JSON file creation which in the future, it has used to show the result of spoofing and I have tried to find a good way to visualize our data. We were divided into small groups and I worked into the machine learning. I have done research about feature selection because of the limited knowledge about the data and to find good attributes for machine learning. Furthermore, I have researched to find a good way to normalize our data because we have big difference range of value for our attribute. I have changed our pom.xml to make an uber jar (a file that contains all files) and a shaded one which splits the source code and dependencies into two different jar files.

In the rest of time, I and Emmanuel have worked together to develop unsupervised learning. I have research and try different clustering algorithms such as K-NN, DBSCAN, Gaussian Mixture Models. We choose K-Means and Gaussian Mixture Models for unsupervised learning because it is supported by Spark MLlib. I have to explore Gaussian Mixture Models in Matlab because Spark MLlib does not have a good library which supports Gaussian Mixture Models. After the midterm presentation, me and Philip have discussion about how HDFS and Spark will be implemented in our new clusters for a benchmark. After that, I have merged Gaussian Mixture Models and K-Means to have the same approach so we could easily do maintenance for unsupervised learning. I have done some discussion with Emmanuel to understand more about K-Means and how we want to merge these two algorithms. In the end, I produced a visualization about the unsupervised learning results to get a better understanding about our clustering algorithms results.

In this project, I have learned a lot of things. I have learned about teamwork, project management, pair programming, communication skills. With SCRUM methodology, I got an experience in project management in the a project. I get more experience and knowledge about Spark, HDFS, and some new machine learning algorithms which I did not know beforehand. Another thing, I have a good experience with code convention and versioning of code. I have learned how to build a clean code and with

GitHub, I learn how to maintain our code with eight other people working together. In conclusion, I have been very happy about this project course because I could work with eight different people with different background. I get a lot of new experience and knowledge in most aspects which would be needed in the real world.

Rahul Setty

The first two weeks was mostly on the research part where all the team members read through Scila description, Apache Spark Java framework (Parsing data, Structured Query Language, Machine Learning etc), financial terms related to Scila, github basics and setting up the environment in local machines through IntelliJ. After that, we divided the work among ourselves. Me and Daniel started to work on the parsing part with two approaches which was mentioned in the Apache Java framework for parsing the Scila data. I worked with Satya on the first approach and he helped me in defining schema for different message types. After me and Daniel completed parsing data with above approaches, we compared how fast each approach is taking to parse data. We observed that the 2nd approach which is parsing data using Java beans with encoders was much faster that 1st approach. We addressed this to Scila and they also suggested to continue with 2nd approach.

Then I decided to work on the supervised part of Machine learning where we have to predict the new data using previous data. I initially started with Support Vector Machines (SVM) which is a classification technique. There were lot of approaches mentioned in Apache Spark Java framework such as Logistic regression, Random forest etc. I tried most of the approaches to check how each approach predicts the new data. Asa also started to work on supervised learning along with me. We also had to parse the data again to convert into SVM (Support Vector Machine) format for it to use the Support Vector Machine function of Apache Spark. Both of us used different approach to parse the data and then we merged both the approaches into one program. Asa continued to work on supervised learning part and I started to work with the Time series model.

The time series model is used to perform forecasting. In this scenario, we are forecasting closing price of the stock from Scila data. Initially, I did research on the ARIMA (Auto-Regressive Integrated Moving Average) model which is a time series model. First, time series data was created from the parsed Scila data. Then the ARIMA model was applied on the time series data. First, I had used a third party module for ARIMA in Java Apache Spark. But the module was unstable and did not produce convincing results. So, I changed the framework and started to work in R programming language to implement the ARIMA model. Therefore, the time series data is generated in Java Apache Spark framework and ARIMA model is implemented on the time series data in R language. The ARIMA model in R gave convincing results though not completely convincing since the data from Scila was a generated data (not original customer data).

It was a good experience for me to work in a group project. Through this project, I got a chance to work with new framework (Apache spark, IntelliJ, Maven, Spring), Git version control and new programming languages such as R programming. I also got to know how the financial sector works especially how stock exchanges work, detect frauds from data etc. and various machine learning techniques such as time series, logistic regression, random forest etc.

Satya Vrat Shukla

At the start of the project, like everyone else, I too spent time researching the various topics we needed to know so that we could begin out work. We spent the first two weeks going through the Scila specifications, and getting to know the Apache Spark framework, familiarizing myself with Hadoop and the basics of git. I spent further time afterwards exploring the financial aspect of the project as I was fairly intrigued with them. The workings of a stock market, the steps required for an order to be placed and converted to a trade, the many ways in which financial fraud is carried out and how the industry is structured and how the regulators go about figuring out and catching fraud. The different types of fraudulent activities that exist and out of those, we were tasked to look more closely at Spoofing.

After the initial weeks, I first helped Rahul in his approach with parsing the Scila data and defining the schema programmatically for different message types. After comparing this approach with the one developed by Daniel, it was decided that the using Java beans with encoders was much faster.

After this, most of my time was spent working with Martin on the spoofing component of the project. Over a number of sprints, we tried different approaches on making a spoofing algorithm that could select the orders suspected of this specific fraud. Our initial few approaches were dependent upon first looking at orders and then making our way to the trade side of the order book. But this led to a number of problems in defining the parameters as suggested by Scila. This was how decided upon our next iteration where we first looked at trades and then backtracked towards orders, narrowing the scope of our filters so that by the end of all their application, we were left with only a few selected orders that would have a high possibility of being spoof orders.

Along the way we had a number of consultation sessions with Scila that that helped in clearing the doubts that came up and gave us the assurance that we we were on the right track, and make the changes where they were needed.

Overall, I have really enjoyed the time working with my smart, hardworking people who were always ready to help around and overall, I think, we had a really good working environment that allowed us to get the best out of this course.

Ludvig Strömberg

At the beginning of the course I spent a lot of time researching topics regarding spoofing and financial markets to get a good general view of the project itself. A lot of time was also needed to learn Apache Spark since I had no prior experience working with it, after getting a better understanding of the software I switched my area of focus to learn more about Spark SQL. Exploring Spark SQL I wrote a variety of queries for testing and understanding purposes, a lot that was later used as reference for further implementations.

At the beginning of the development I helped to create the algorithm for spoofing detection, both trying to develop the algorithm in a logical sense but also creating functionality in Spark to try and achieve what was needed in regard of the specification.

After working a bit with spoofing I focused more on Machine Learning and in particular the external Deep Learning 4 Java library, I wrote the transformer that converts our Spark data to make it ready-to-use with DL4J. I also tried some test implementation of a neural network in DL4J but after realizing the provided dataset being totally randomized and not giving any interesting results I shifted focus.

I wrote the bash script to generate more sample data, with the purpose to have a bigger set of data for benchmarking purposes. I also wrote a bash script for running our benchmarking.

Other than the previously mentioned areas I worked with a little bit of everything and tried to help wherever it was needed, for example trying to brainstorm with someone who perhaps got stuck working with some problem. I also bought plenty of fika for the office :)

Overall the course was a great experience on working within a larger project, compared to previous courses I taken it felt like this course offered an environment more similar to working outside the world of academia. I would say the main new skills I have obtained are within Apache Spark and working with SCRUM. The course definitely helped with practically improving my SQL, Java and GitHub skills.