**Introduction**

The basic idea of the project is to develop a tool, which would deploy one or more topologies on Spark on the Department of Computing(DoC) cloud automatically. Moreover, the tool should also allow the users to evaluate the performance of different topologies, and analysis various utilisation data (CPU Usage, Memory Usage, Response Time etc.) conveniently and directly.

**Methods**

We will choose **Scrum** method.

The reasons behind choosing the Scrum method are numerous:

* Using the Scrum there is a better organization between the team. This is because we have three different main roles and everyone knows clearly what to do. This is very important for our team since we have to undertake a big project in a short time and we need to have a good organization among the team.
* After each Scrum sprint we can release some working features of the product. This help us to have something that works in a short time. Also, coupled with regular meetings with the supervisor, we can have a better understanding of specifications with a clear feedback from him because he can see something that actually works and redirect us on the right path. Therefore, we can detect issues early which can be resolved in speedily. Moreover, the release of each feature can assist us to manage a predictable schedule.
* Further, following the scrum method all the members can be more productive. This is due to frequent meetings that make it possible to check everyone’s productivity.
* In addition, through this method we are being able to see the incrementation of the progress in work due to daily meetings. Thus leading us to have a clear visibility of the project advancement.
* We prefer to work individually. It is unlikely for us to work together everyday, as all of us living in different accommodations. Working individually and then have a daily meeting makes things easier.
* We want a methodology that pushes everyone to be more committed. In SCRUM, the whole members have to work as one to provide well defined schedules and activities. As a result, everyone can be more motivated.
* Scrum helps everyone in the team to have an overall view of the project. As a matter of fact, it provides a product backlog, with all the accumulated information of the project, to each member of the team.
* Scrum eases the workload of the team members regarding scheduling and managing Therefore, that leads us to be more focused on the quality of the product which means a code with less bugs and not poorly-written.

The methodology that we are going to follow:

**Roles and Responsibilities**

|  |  |  |
| --- | --- | --- |
| **Name** | **Role (Scrum & Belbin’s Team Roles)** | **Responsibility** |
| Giuliano Casale | Supervisor | Supervise the project |
| Pooyan Jamshidi | Technical Advisor & Product Owner | Provide project descriptions. |
| Haoran Zhang | Coordinator & Scrum Master | Guide the team to tackle the objectives and manage the development process of the project. |
| Cong Lu | Monitor-Evaluator | Analyze and evaluate new ideas that the others come up with. |
| Yao Liu | Resource Investigator | Evaluate the background and clearly understand the requirements of the project. |
| Fahd Abdeljallal | Shaper | Push the team to consider a lot of possibilities in order to find the best solution. |
| Jiaxin Qiu | Plant | Come up with new ideas and approaches. |
| Chadjiminas Ioannis | Complete-Finisher | Check that the project has been implemented carefully without any errors or omissions within time-limit. |

**Planning**

1. Project Plan

The project that we are going to undertake requires us to carefully understand how to design and implement it. Hence, initially we did lot of research regarding the Apache SPARK. Following this, after regular team meetings we have discussed the steps that we need to follow in order to finish the project successfully within the deadline. Through the team meetings we have decided also the programming languages and high-level tools that we are going to use. Further, we have assigned for each team member different roles so that everyone clearly understands his position and responsibility in the team. In addition, we have coordinated meetings with our supervisor and technical advisor to obtain feedback.

**Programming Languages & High Level Tools**

* Apache Spark:
* Java:
* Gitlab:
* Trello:

**Project Details**

The functional requirements represented below that must be achieved in the final application.

|  |  |
| --- | --- |
| Requirement | Priority |
| Deploy a topology automatically | Must |
| Change configuration parameters of Spark | Must |
| Provide monitoring suite | Must |
| Store performance data | Must |
| Un-deploy the cluster | Must |
| Clean up the cluster | Must |
| Get utilization data | Should |
| Integrated data sources | Should |
| Integrate some existing benchmark topologies | Should |

The intuition is that we could achieve this by doing things below:

1. Firstly, create clusters on the DoC cloud, and deploy a topology manually, such that we could get familiar with the installation and configuration of Spark rapidly.

2. Read the documentation of Spark, and try to figure out which APIs we may need for the automatic deployment. E.g. The API for nodes configuration.

3. Develop a tool which would deploy the topologies automatically by sending a configuration file to a master machine, and the master machine would deploy the slave machines automatically according to the configuration file (Figure 1), or sending the configuration files to each machine in the cluster, and all machines deploy itself based on its configuration file (Figure 2). We are not sure about which method is better and feasible for now, so we just make an assumption here. Moreover, the tool should also provide a REST API for the users to modify the configuration parameters of the Spark.

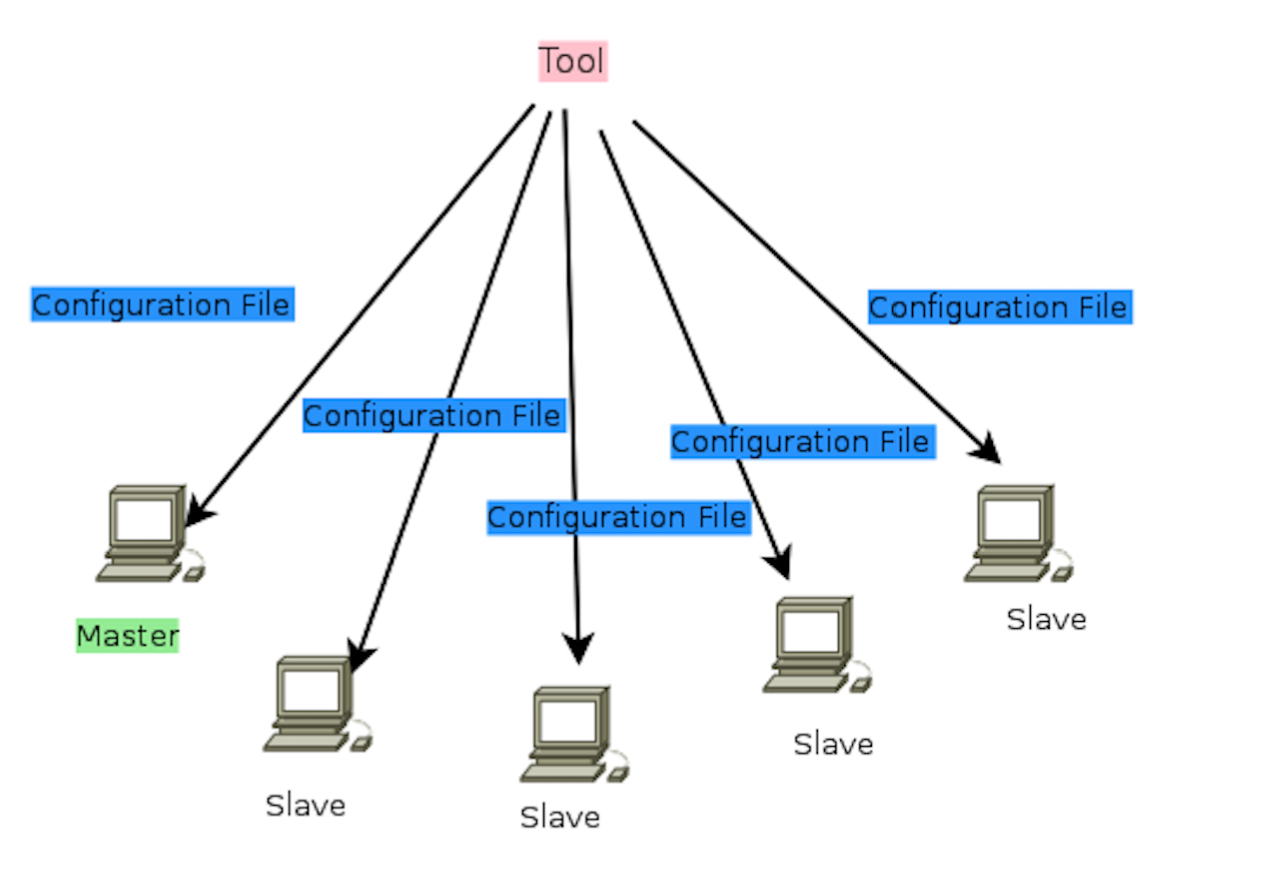
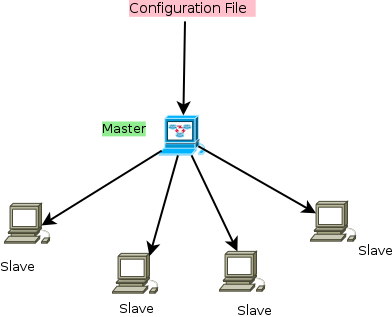


Figure 1 Figure 2

4. The tool should also provide a view for the users to monitor the statues of the whole cluster and each machine within the cluster. As the configuration of the Spark may be separated to different phrases, so the tool should also provide a way to reflect the system performance (E.g. CPU Usage, Memory Usage, Response Time) in each stage.

5. We should also provide a method to illustrate the performance the different topologies. E.g. Store the performance data to CVS file, or create a GUI to provide some graphs and charts for the users to evaluate and analysis the information they want directly.

6. Integrated the tool by combining each part together.

7. Testing with existing benchmark topologies to indicate the usability of the tool.

--------------------------------------------------------------------------------------

**2. Project management (Including risk management)**

In order to perform the whole development progress smoothly, regular evaluation is necessary. Since Scrum has been chosen as the development method, regular meeting will be hold and the evaluation of works will be reflected. Scrum Master is in charge of making sure that the process is in right track.

However, there are also some potential risks during the development progress. In order to avoid them, some backup and alternative plans should be prepared in advance. If things go wrong, the corresponding arrangements can be applied immediately.

1. Technology Risk

All data and documentation related to the project should be backup with a removable disk during the software development process. In case of crash of servers or failure of connection, it may cause loss of data, however data can be easily recovered and redistributed through backup.

Likelihood: Unlikely

Severity: Tolerable

1. Human Risk

In case that team members can be ill, they can not attend a meeting and achieve the latest information of the project. Therefore, in every meeting, there needs to record any arrangement and discussing views to inform other members.

Likelihood: Likely

Severity: Tolerable

1. Schedule Risk

n case that any time period in Gantt Chart can be not adequate for the corresponding workload, it can delay delivery of output. Hence, in every beginning of plan, it needs to arrange more flexible time in each period and update process information in meetings.

Likelihood: Likely

Severity: Serious