

DongshengCao_Proposal

by Dongsheng Cao

FILE	12039_DONGSHENG_CAO_DONGSHENGCAO_PROPOSAL_93935_157193776.DOCX (748.05K)		
TIME SUBMITTED	08-MAY-2019 09:57 PM (UTC+0100)	WORD COUNT	3998
SUBMISSION ID	106697786	CHARACTER COUNT	25325

Spark-based Heatmap Visualization
Research Proposal
Dongsheng Cao

Table of Contents

Figures.....	3
1. Introduction	4
2. Background and to the Proposed Research.....	4
2.1 Background of visualization tools	4
2.2 Background of my study fits into the project	5
3. Aims, objectives and the expected outcomes of the project	5
4. Literature review.....	5
4.1 Background	5
4.2 Related literatures	6
5. Research Methods	6
5.1 Philosophy.....	7
5.2 Strategy	7
5.3 Approaches	7
5.3.1 Document Review	7
5.3.2 Software Development	8
5.3.3 Case Study.....	8
5.4 Risks & Limitations	8
6. Development Methods	8
6.1 Software Development Methodology	8
7. Resources & Planning.....	9
7.1 Resources Required	9
7.2 Gantt Chart.....	10
7.3 Some special notes	10
References	11
Bibliography	12

Figures

1. Figure 1: Waterfall model.....	9
2. Figure 2: Prototyping Methodology.....	9
3. Figure 3: Gantt Chart	10
4. Figure 4: Gantt Chart	10

1. Introduction

In this Big Data era, facing to the data scale from TB to the PB, EB and even the ZB level, the experts in this area are meeting a variety of limits and obstacles in terms of data analysis. Besides, the 3 'V', Volume, Variety and Velocity, is seen as the most outstanding characteristic of Big Data, which aims to combine and fuse the various original data together to meet the potential value of the data. In details, data mining and visualization [1] in this situation properly provide the platform and the technique to make the data more understandable. In details, the data mining is the method to explain the big data at the bottom while the visualization is the way to explain the data at the display layer.

Specifically, regarding to the situations that facing the large dataset to display whole dataset on the web, the current sort of data visualization tools or methods via web browsers cannot perform quite well, this project is to develop a visualization framework based on the Spark which is a general big data platform can solve core problems such as batch processing, interactive query and the stream computing in large data computing and the main idea came from Zhang Fan team's literature [2]. In addition, and for instance, this project is to develop visualization framework taking heat map as the example with the Spark platform and the LOD (levels of details) method.

In details of this project, on the first place, parallel computing the tile-based algorithm for layered dataset on the Spark platform aims to generate transparent geographic heat-map images which will be used in the next steps to overlay each tile of the map on each layer. Secondly, I will store these images generated in the first step on the HDFS (Hadoop Distributed File System). Finally, using these computed results, I will design and develop a web application to provide various kinds of spatial-temporal analysis services with the geographic information of the dataset and some other kinds of data analysis of the dataset by the other visualization ways such as pie chart, bar chart and so on.

2. Background and to the Proposed Research

2.1 Background of visualization tools

With the increasing requirements for web visualization, there are some web visualization tools like D3, heatmap.js, the Matplotlib and Boken based on Python and so on. D3 manipulates the documents of data, which is a JavaScript library to help people to make data to life by CSS, HTML and SVG (Scalable Vector Graphics). What is more, D3 actually has a variety of advantages but it is not suitable facing the data above 1 billion points [3]. Similarly, the heatmap.js using canvas to draw data also does not fit the capacity of large volume of data. In addition, the Echarts from Baidu is similar to the D3 as well.

Heatmap is a kind of visual representation of a large number of data points with a special feature in a given area, which means a specific position in the map of a certain point is used to map a certain geographic object. Besides, the concentration parameter of this point represents the value of the elements, features or attributes of the geographic object and then linearly stretching. Therefore, the heat map is the method to explain feature with the varying value scope of the third dimensional information of the data [4]. In the recent years, heatmap has been used to analyse various kinds of data. For example, Scott's team using heat-map to analyse the multidimensional gun/homicide data [5]; Xiao Chen and Genke Yang [6] introduced a limb detection heatmaps to represent the body joint pairs association.

2.2 Background of my study fits into the project

- Have learnt Web Technologies for eCommerce by Gaynor Thomas.
- Have learnt Distributed Web Application by Tim.
- Have learnt some knowledge of data analysis by Mabrouka.
- Have learnt some Knowledge of data visualization by Nik.
- Have been learning the knowledge of development of big data.

3. Aims, objectives and the expected outcomes of the project

The main aim of my project is to justify whether this idea of data visualization is effective and efficient that using Spark parallel computing platform to provide computing capacity to the big data visualization via Web browser, and the main work is to develop a web visualization framework based on Spark to evaluate and test.

There are some objectives will need to be achieved in order to realise the aim of project:

- The heatmap visualization via Web browsers realizing 5 layers of tile map.
- To do a series of data analysis of the dataset displaying on the web.
- To realise the function in terms of real-time aspect.
- To achieve some optimized functions such as using distributed database (Hbase), Ajax to dynamically transfer data of map and so on.
- To test the performance of the data visualization framework with Spark this project will be achieving and compared with the common data visualization tools at the same situations.
- To test the optimized functions whether they are beneficial to the frame.

Finally, at the end of project, it is expected to realise a well-performed and faster web visualization frame with Spark, a variety of data analysis dataset displaying via web browser and the optimised functions can achieve and will be tested well. In details, in this case of my project, I expected the heatmap visualization of the dataset can perform well in terms of speed while zooming in and zooming out on the map and the comfortable exhibition of the tile map, the data analysis to display more information came from dataset displayed more understandable and the optimised functions can improve the framework more efficient.

4. Literature review

4.1 Background

Michael Friendly thought [7], the purpose of data visualization is actually to visually display data, such as converting the amount of data that takes hours or more to be summarized into an index that can be read at a glance; by adding, subtracting, multiplying and dividing, and weighing and calculating the difference between the two groups of data, the colour sensitivity, length and size of the image can be compared; data visualization is a powerful tool for communicating complex information. Besides, visualizing information, our brains can better capture and preserve effective information and increase the impression of information. Data visualization is treated as an equivalent thing of visual communication, which includes the composition and study of the visual displaying of the data.

Stephen Few thought [8], in order to clearly and efficiently link the information of data, there are many kinds of tools in terms of data visualization such as plots, information graphics,

statistical graphics and others. Besides, it is better to be encoded for the numerical data using lines, dots or bars in order to visually communicate with a quantitative dataset.

In the opinion of Manuela Aparicio and Carlos J. Costa [9], 'Data visualization is both an art and a science', it is treated as a section of descriptive statistics somewhat and also as a fundamental theory of development tools developed by others. And Nikos Bikaks addressed [10], the increasing number of data created by the Internet and the a number and variety of sensors in the various environments are seen to as 'big data' and the processing, analysing and communicating with these data are generating various ethical and analytical challenges for the data visualization.

4.2 Related literatures

This section is to introduce some sorts of literatures relating to the improvements of data visualization in the various aspects.

An advanced aggregate computation [11] was introduced by Xinxiao li, Akira Kuroda, etc. In details, they aimed to deal with the issues in terms of three aspects. First one is mapping the large dataset into a visualization-based small dataset in order to perform efficiently and keep operability of the dataset. Second point is to minimise the scale of visual original data to the aggregated data in order to perform the visibility of the important information of the dataset but use the tiny dataset. In addition, the final point is to use the third-party tools like D3 that D3-based rendering implementation can improve the performance of the continuous visualizations.

An improved method on data visualization [12] regarding to the large-scale terrain surface was proposed by Huaqing He, Chong Wang, etc. Within their approach, based on the ROAM algorithm to improve its performance, they mainly focus on the two aspects: one is the LOD control which is the view-dependent thing for the level of detail usually be used in the terms of map area such as heat map for the geographic information on the map and in my project and the out-of core data paging.

Shu Gao and Ding-Fang Chen proposed a distributed visualization applying the web service technology [13]. In their paper, they discussed how dataflow visualization could improve into the way of grid computing regarding to the web service techniques used to design the environments of distributed visualization. And the other focuses of their paper are the architecture and implement of a web service-oriented prototype of distributed visualization

Zhang Fan and Yuan Zhaokang [2] proposed an idea that combining the Spark platform to provide parallel computing capacity and the HDFS to provide storage resource with the data visualization via web browser in order to improve the performance of data visualization on the web. Especially, my project's main idea certainly comes from their paper and my main aim is to develop a web application with this data visualization framework to justify whether it is available and efficient.

5. Research Methods

In this chapter, I will discuss what sorts of research methods I will adopt in my project, including the Philosophy, approaches and the strategies. In addition, I will also discuss the risks I may meet in the process of project and how to properly deal with and solve these difficulties. The data evaluation will be introduced as well in this chapter.

5.1 Philosophy

According to Sinha [14], the main point of positivism is the experience, which means the conclusion must come from objective situations and the external things through the observation or sensory experience and so on. Because, although each individual with different education background, there is not much difference of the concepts they use to validate sensory experience. In a word, the purpose of positivism is to create the objectivity of the knowledge. By contrast, phenomenology is certainly conscious point that knowledge is seen as social phenomenon and can only be treated subjectively. In details, the collected data of it is subjective in order to donate to the ideas, feeling, thoughts and some other kinds of subjective points.

For my project, I will adopt positivist philosophy method to justify whether this idea of data visualization is effective and efficient that using Spark parallel computing platform to provide computing capacity to the big data visualization via Web browser.

5.2 Strategy

According to lecture's PowerPoint, in the applied computing area for the short project, there are 5 kinds of common-used strategies such as experiment, software development, document review, case study and the survey.

There are main three kinds of projects such as experimental projects, development projects and the survey project. In details, experimental projects usually use experiment, software development, document review. Development projects always use software development, document review and the case study. Besides, survey projects usually use document review, case study and the survey. My project could subject to the kind of development project, because the main works of my project are related to areas of developing web application.

Therefore, for my project, I will choose the strategies of document review, software development and the case study, which is quite common strategies used by development project.

5.3 Approaches

There are two common approaches of research methods including deductive and inductive. The deductive usually consists of theory, hypothesis and the research strategies to test this hypothesis. On the other hand, the inductive is a process of collecting data, analysing this data and through them to develop the theory.

For instance, I will adopt the deductive method for my project because my project is a developing project which aims to justify my new piece of software performs better. In details, the hypothesis of my project is whether the new web visualization frame I will develop based on the idea to work wit Spark platform can get a better performance than the basic one without the Spark platform. In addition, I will adopt the strategies of software development, document review and the case study to test this hypothesis.

5.3.1 Document Review

Document review will be adopted for my project, because through the a series of literatures, I can identify the main ideas to improve the performance of data visualization, identify the main theories in this area and confirm some steps can realise in practice such as using the Spark platform to provide parallel computing capacity, using HDFS to provide storage and so on. The details are introduced at the chapter 4 literature review.

5.3.2 Software Development

Software development will also be adopted for my project, because through the development software, I can prepare basic conditions to the next steps to justify my hypothesis. The advantage of this strategy is that it can directly to show the results which can help me test with my data. However, there is still a problem exists that in the software development stage, it is possible to not achieve some expected functions of the software. Regarding to this risk, I will carry out a lot of literatures reviewing to make sure whether these function can achieve in the current situations.

There is a main work in my project that is to develop a web application of data visualization to realise an idea that combine the Spark platform with the application in order to improve the performance. What is more, the works of software development are as following: establishing and configuring the Spark computing platform and the HDFS (Hadoop Distributed File System), establishing the web application including the HTML pages, CSS files, JavaScript files, Servers for the Web application and so on.

5.3.3 Case Study

Case study will be adopted for my project as well. In details, because I can use this way to test my hypothesis. Specifically, I will use a heatmap case of the traffic data with the geographic information to test whether the new piece of software is better. The main focuses of this case are to establishing a tile map with several layers in order to display more detailing information on the map when zooming in and zooming out, the principle of which is using the painted transparent heatmap images of the dataset to overlay each tile of the map at each the layer.

In details, this task includes several subtasks such as transforming raw data to the available data, drawing data with heatmap format and generating the transparent heatmap images classified by the index of tile at each layer, storing these images on the HDFS and finally, overlaying the transparent heatmap images on the map and showing the whole map via web browser, the computing subtasks of which are on the Spark platform.

5.4 Risks & Limitations

There are several issues or risks I maybe encounter in the process of my project and the solution of these issues

- Cannot complete the work on time – strictly set up and carry out the review meeting weekly with my mates to review our works respectively and report works weekly to supervisor.
- Struggle to realize some functions – ask my supervisor, other teachers and mates for help and check some relating forums.
- Don not have enough data to test – make some more complex iterative process to increase the amount of calculation and in the meanwhile to find some relating data to mix properly.

6. Development Methods

6.1 Software Development Methodology

In the software development area, there are a variety of popular development methodology [15] such as waterfall development methodology, prototyping methodology, agile development

methodology, DevOps development methodology and the rapid application development methodology.

The waterfall Model [16] is the quite traditional method to explain the basic development process: requirement gathering & analysis, system design, implement, testing, deployment/installation of system, maintenance. In addition, there are some obvious advantages such as easy to understand, saving time and simple to handle, etc. However, there also are some serious problem of it such as only matching precise requirement and not excellent for the ongoing and long project.

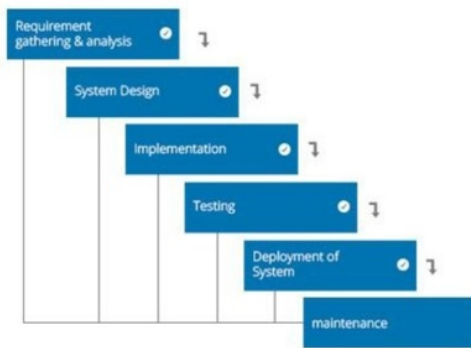


Figure 1: Waterfall model

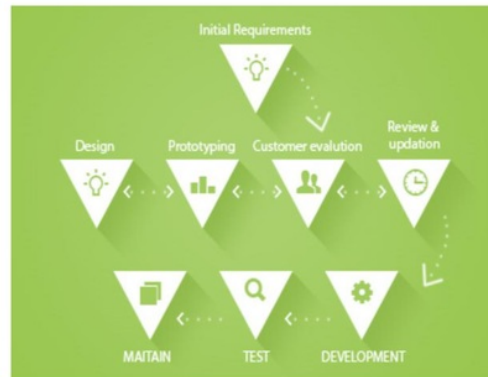


Figure 2: Prototyping Methodology

The prototyping methodology [16] is a specialized procedure of software development and it is to achieve the fundamental functions of the software making some samples of the resolution. In details, the aim of it is to make some important changes before establishing the final solution. In addition, it is fact that the prototyping methodology tends to cope with a set of various problems with the waterfall method. There are some advantages of the prototyping methodology: giving clear idea about the functions of the software, reducing the risks of failure of the functions and assisting well in terms of requirement and the whole analysis.

Through the comparison of the waterfall and prototyping development methods, I will adopt prototyping methodology for my project, because my project is to test my hypothesis that means there will be some changes in the developing process and using prototyping methodology can reduce some failure risks of some functions difficult to achieve, which properly meet the requirement of my project.

7. Resources & Planning

7.1 Resources Required

There are a set of hardware and software resources and some code libraries to assist achieving some functions needed for my project:

- Three computers – will be used to make up of a Spark cluster.
- Microsoft Word, Microsoft Project, Microsoft Excel, Microsoft PowerPoint – will be used to planning, recording the requirements, scheduling and analysing data, etc.
- Hadoop, Spark platform – will be used to establish, configure and managing the cluster consists of 3 nodes.

- HTML, CSS, JavaScript, OpenLayers, GeoServer, Echart, Scala – will be used to establish the functions of data visualization
- Related tutorials – will support me to program.

7.2 Gantt Chart

According to prototyping methodology, the Gantt Chart has been built as following:

	Establish and configure Spark cluster	2 wks	Mon 20/05/19	Fri 31/05/19	
	Data collection	1 wk	Mon 03/06/19	Fri 07/06/19	1
	Data cleaning	1 wk	Mon 10/06/19	Fri 14/06/19	2
	draw transparent heatmap images	2 wks	Mon 17/06/19	Fri 28/06/19	3
	Do some kinds of data analysis displaying on the web page	4 days	Mon 01/07/19	Thu 04/07/19	4
	Establish web application with heatmap visualization	2 wks	Fri 05/07/19	Thu 18/07/19	5
	Other optimized functions	2 wks	Fri 19/07/19	Thu 01/08/19	6
	Function reviews	34 days	Mon 17/06/19	Thu 01/08/19	
	Writing report	20 days	Fri 02/08/19	Thu 29/08/19	7

Figure 3: Gantt Chart

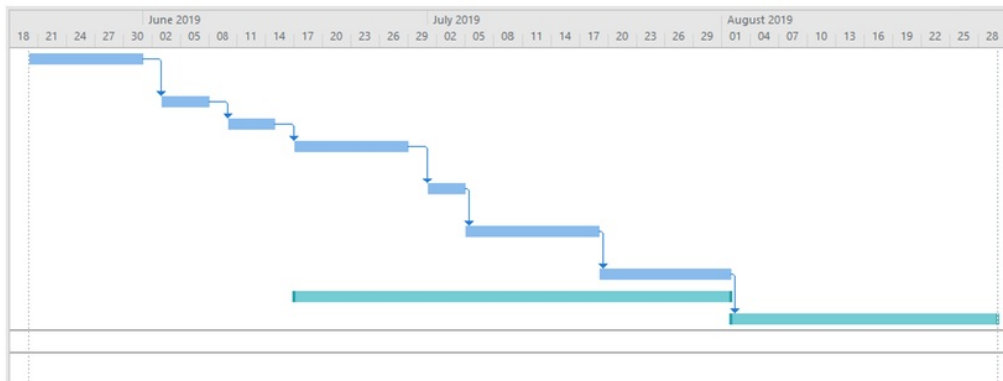


Figure 4: Gantt Chart

7.3 Some special notes

The first task, establishing and configuring Spark cluster, last 2 weeks because it is not easy to establish a Spark without any problems.

The task, Other optimized functions, include using distributed database (Hbase), Ajax to dynamically transfer data of map and so on.

The task, function reviews, is quite important part in the process of my project, which is carried out at the very Friday from 17th of Jun to the 1st of August. The aim of it is to make sure the works done in the last week and check whether the functions implemented well or whether any function need to change and so on.

References

- [1] S. Holtz, G. Valle, J. Howard and P. Morreale, "Visualization and pattern identification in large scale time series data," in *2011 IEEE Symposium on Large Data Analysis and Visualization*, Providence, RI, USA, 2011.
- [2] F. Zhang, Z. Yuan, F. Xiao, K. You and Z. Wang, "Research on Heatmap for Big Data Based on Spark," *Journal of Computer-Aided Design & Computer Graphics*, vol. 28, no. 11, 2016.
- [3] B. M, O. V and H. J, "D3: data-driven documents[J]," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301-2309, 2011.
- [4] Z. Yi, G. Bin and L. Qiming, "A Method of Heatmap for GIS 3 Dimension Data," *Computer Programming Skills and Maintenance*, vol. 22, pp. 120-122, 2011.
- [5] S. D. Rothenberger, J. E. Wenskovitch and G. Marai, "Pexel and Heatmap visual analysis of multidimensional gun/homicide data," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Providence, RI, USA, 2011.
- [6] X. Chen and G. Yang, "Multi-Person Pose Estimation with LIMB Detection Heatmaps," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018.
- [7] M. Friendly, "Milestones in the history of thematic cartography, statistical graphics, and data visualization," 2008.
- [8] S. Few, "Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message," 2004.
- [9] M. Aparicio and C. J. Costa, "Data visualization," in *Communication Design Quarterly Review*, 2014, pp. 7-11.
- [10] N. Bikaks, "Big Data Visualization Tools," in *Encyclopedia of Big Data Technologies*, Springer 2018, 2018.
- [11] X. Li, A. Kuroda, H. Matsuzaki and N. Nakajima, "Advanced aggregate computation for large data visualization," in *2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV)*, Chicago, IL, USA, 2015.
- [12] H. He, C. Wang and H. Liu, "An Improved Approach on Visualization of Large-Scale Terrain Surface," in *Computer Graphics, Imaging and Visualisation (CGIV 2007)*, Bangkok, Thailand, 2007.
- [13] S. Gao and D.-F. Chen, "Applying Web Service Technology in Distributed Visualization," in *2007 International Conference on Machine Learning and Cybernetics*, Hong Kong, China, 2007.
- [14] D. Sinha, "Philosophy and Phenomenological Research," *International Phenomenological Society*, pp. 562-577, 1963.

- [15] S. E. Team, "Top 4 software development methodologies," Synopsys, 28 March 2017. [Online]. Available: <https://www.synopsys.com/blogs/software-security/top-4-software-development-methodologies/>.
- [16] Acodez, "12 Best Software Development Methodologies with Pros and Cons," Acodez , 1 Jun 2018. [Online]. Available: <https://acodez.in/12-best-software-development-methodologies-pros-cons/#Pros-2>.

Bibliography

- [1] S. Holtz, G. Valle, J. Howard and P. Morreale, "Visualization and pattern identification in large scale time series data," in *2011 IEEE Symposium on Large Data Analysis and Visualization*, Providence, RI, USA, 2011.
- [2] F. Zhang, Z. Yuan, F. Xiao, K. You and Z. Wang, "Research on Heatmap for Big Data Based on Spark," *Journal of Computer-Aided Design & Computer Graphics*, vol. 28, no. 11, 2016.
- [3] B. M, O. V and H. J, "D3: data-driven documents[J]," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301-2309, 2011.
- [4] Z. Yi, G. Bin and L. Qiming, "A Method of Heatmap for GIS 3 Dimension Data," *Computer Programming Skills and Maintenance*, vol. 22, pp. 120-122, 2011.
- [5] S. Few, "Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message," 2004.
- [6] M. Aparicio and C. J. Costa, "Data visualization," in *Communication Design Quarterly Review*, 2014, pp. 7-11.
- [7] N. Bikaks, "Big Data Visualization Tools," in *Encyclopedia of Big Data Technologies*, Springer 2018, 2018.
- [8] S. Gao and D.-F. Chen, "Applying Web Service Technology in Distributed Visualization," in 2007 International Conference on Machine Learning and Cybernetics, Hong Kong, China, 2007.
- [9] RL Glass, I Vessey, V Ramesh - Information and Software technology, 2002 - Elsevier
- [10] Zaharia, Matei; Chowdhury, Mosharaf; Das, Tathagata; Dave, Ankur; Ma,, Justin; McCauley, Murphy; J., Michael; Shenker, Scott; Stoica, Ion (2010). Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing (PDF). USENIX Symp. Networked Systems Design and Implementation.
- [11] Wilkinson, Leland; Friendly, Michael (May 2009). "The History of the Cluster Heat Map". *The American Statistician*. 63 (2): 179–184.

- [12] Perrot, A.; Bourqui, R.; Hanusse, N.; Lalanne, F.; Auber, D (2015). *Large interactive visualization of density functions on big data infrastructure*. *IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV)*, 2015. pp. 99–106.
- [13] "ol/layer/Heatmap~Heatmap". *OpenLayers*. Retrieved 2019-01-01.
- [14] Manuela Aparicio and Carlos J. Costa (November 2014). "Data visualization". *Communication Design Quarterly Review*.
- [15] "Spark 2.2.0 Quick Start". *apache.org*. 2017-07-11. Retrieved 2017-10-19.

DongshengCao_Proposal

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

Instructor

65/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

GRADING FORM: NEW ASSESSMENT

DONGSHENG CAO

65

ENGAGEMENT WITH LITERATURE

-  Relevant background literature has been reviewed to highlight some of the methods used to address performance issues for data visualisations of large datasets on the web. It would, however, have been nice to see some of the results from these studies discussed as well.

11

KNOWLEDGE AND UNDERSTANDING

-  Good introduction to the project identifying its relevance. Generally, a well-defined project aim and set of objectives.

10

COGNITIVE, INTELLECTUAL SKILLS

-  An appropriate development methodology has been identified and appropriate software and hardware resources needed for the project have been identified.

10

PRACTICAL SKILLS

-  Good discussion of proposed research methodology with justifications for the chosen methods, but the student needs to think of how he is going to determine if his proposed solution performs better as stated in his hypothesis or think of alternative methods of analysis. The proposal includes a

28.5

project plan that has been generally well thought out for the implementation, but the student needs to revisit the time allocated to the project report. For example, the student will need to conduct a full project literature review early on in the project that will be used to inform the rest of the project work.

TRANSFERABLE SKILLS



The document is satisfactory, but work at improving your academic writing skills.

5.5