**Project Proposal Report**

**Student Details**

Full Name: Braznell, James

Student Number: 1007022

Award/Course: BSc (Hons) Computer Science

University email address: [J.Braznell@wlv.ac.uk](mailto:J.Braznell@wlv.ac.uk)

Supervisor Full Name: Sherin Nassa

Date of Submission: 21st January 2013

**Contents:**

1. Student details (page 1)
2. Contents page (page 2)
3. Project details (page 3-4)
4. Project proposal (pages 5-11)
5. Additional information (pages 12-22)

**Project Details**

Project Title:

What database technologies are emerging to handle incredibly large size of data?

British Computer Society:

It is my intention that my project does not meet the requirements of the BCS

Academic Question:

Which of the emerging database technologies best handle the problem of collecting, storing and analysing incredibly large size of data?

Aims:

To research the emerging database technologies and from this create an artefact that can be tested against other current systems to determine what technologies best handle incredibly large size of data.

To create a report based on testing of the artefact that details the opportunities that can be gained from exploiting big data and what type of businesses can benefit from investing in a system capable of handling big data.

To research the technologies that are currently in development, which have yet to be fully adopted and from this create a prediction of where the developments in the handling of big data is heading.

Objectives:

Research big data origins and predictions of future solutions

Research the Hardware specifications needed to efficiently handle big data

Research the emerging database systems for a data storage system which can handle incredibly large size of data

Research the data that is most useful to a business looking to exploit big data

Research how this data can help a business make better decisions

Research the emerging applications that can handle the analysis of incredibly large size of data

Research the current benchmarks for systems that handle incredibly large size of data

Implement a data storage system capable of handling incredibly large size of data

Create a substantial amount of test data that can be used to assess the efficiency of the artefact

Test the data storage system with the newly acquired test data for speed and efficiency

Employ a system to analyse this data in a graphical representation

Test the end data collected with targeted queries

From the results generated from testing and analysing the data evaluate the artefact

Report on the database technologies currently in development and create a prediction of where the developments in the handling of big data are heading

Give conclusions of the project and critically evaluate all aspects

Artefact:

Based on insights gained from research into the issue of big data, I will choose systems that I believe will work best together in order to create a design that I will then populate with test data. This will then be tested to show the efficiency of the design, I will compare these results to benchmarks set for these systems for handling incredibly large size of data. During the development of this artefact the design will be updated on a regular basis in order to improve the efficiency of the finished artefact and therefore create a better design for a solution to big data.

**Project Proposal**

**Introduction:**

This project will show the best design needed to collect, store and analyse what is known as ‘big data’, big data is information from a variety of different sources such as web blogs, location data, network probes, radio frequency identification feeds, sensor data, social media feeds etc. information that a business would not normally be interested in much less have the capabilities to use in any meaningful way until quite recently. Now there are several new technologies that are emerging to handle such a large size of data such as No SQL databases, oracles Exadata, hadoop, map reduce and many more. There are many benefits to an organisation which chooses to invest in a system that can make use of big data, retailers for example can use this data to micro segment their customer base and use sentiment analysis give targeted offers via social media feeds and customer data thereby delivering better revenue and margins. Financial businesses can use locational data, social media and spend patterns to identify fraudulent transaction attempts to make a decision when to stop a credit card. Therefore this will give businesses that choose to use big data further information which keeps them more informed and helps them to make better decisions. I believe that most if not all businesses can benefit from such information and that is why I chose this project, I will research into the existing/emerging technologies that handle big data, create a design that best handles such incredibly large size of data, implement this design and test it to show if it can handle big data. In doing so I will answer my academic question to show which of the emerging database technologies best handle big data by insight gained from research as well as comparing and contrasting these technologies.

**Literature Survey**

Arneil (2012). This source is quite useful as it gives a consumers perspective on Exadata, a summary of exactly what Exadata is, their experiences from the set up, it also goes into detail on the technical specifications of Oracles newly released Exadata X3 which improves greatly on previous versions and finally sheds light on the hardware and software maintenance needed to keep Exadata running. The source provides a good amount of information and does not come directly from Oracle so it is less likely to be biased about Exadata advantages and disadvantages, I will use this information when researching emerging database systems, it will be useful to compare them to Oracle's system for handling big data, this will then be incorporated into the report.

# Carr (2012). This source details the history of big data, claiming it started in 2007 with Google's Map Reduce framework being integrated into Apache Hadoop, it goes on to illustrate how Apache Hadoop and other big data solutions has led to commercialisation of the big data trend. It goes on to list these big data solutions such as Hive, Zookeeper etc, showing that Apache Hadoop is part of a large group of innovations in big data solutions. It also gives its prediction of what is next for big data and how this could be quite different from the open source solutions currently available. The source provides a decent amount of information on big data, the current solutions and its view of the future innovations, therefore this is a useful source that will help further my research and ultimately help improve the artefact, it will be included in the report for a more in depth evaluation.

Clegg (2012). This source outlines what the market is using to handle their big data needs, such NoSQL, Hadoop etc and shows exactly what kind of uses they are popular for, such as on-line big data workloads, batch processing. It also shows why RDBMS are not able to handle the issue of big data and how a good analytics program will help to sift through big data to find exactly what is needed. The source provides a decent amount of information about the big data solutions currently on the market although it could be considered biased by advertising its own analytics program, however I believe it will still be useful and therefore I shall use the source when researching the emerging database systems that can handle big data, this will then be included in the report.

# Howard et al (2012). This source explains what Big data is and how it should be used, it goes on to show how a graphical representation of such data can help to see exactly what you are getting from the data you have collected and stored. This is useful as it provides me with background information on big data and demonstrates how a graphical representation of my artefacts results could help understand the outcome when testing my artefact. The source provides good background information on Big data and a demonstration of analytics tools, therefore this is a very useful source and I will use it to help further my research and will be included in the report.

# Kutrovsky (2012). This source shows how data moves in Exadata in order to get the end result, it also details Parallel query concepts such as the big sort, big group by and a big join which allows for the querying of very large tables in real-time, this is achieved by using a flow-through system of handling the data. This is useful as it details how querying has changed from traditional RDBMS to allow for the increased performance requirements as well as the underlying concepts used to make this a reality. The source is quite useful as it goes into detail on how Exadata functions in order to handle big data and so I will use it in the report when evaluating my artefact.

# Logan et al (2012). This source details the difference when working on an Exadata machine compared to a traditional RDBMS, it discusses how Exadata is tuned to get the best performance out of its hardware, what should be considered when planning for backup's, it discusses the internals of the Exadata machine as well as detailing functions that help with compression, performance and storage. This is useful as it shows how Oracle is providing innovative improvements to hardware and how this differs to traditional RDBMS.

# Powell (2011). This source is useful as it shows some of the most talked about topics in the database industry, such as big data analytics which allows for more in depth and faster analysis of data. Also grid computing which is a form of distributed computing where many grids loosely come together to perform large tasks. It also talks about how Oracles Exadata system is spearheading the innovations directed towards big data solutions into one system providing the performance needed to make use of big data. This is useful because it shows what big data is, it details the tools being crafted to handle such large scale data and how Oracles Exadata is dealing with the issue of big data. Therefore I will use this when researching database systems to handle big data and it will be included in the report.

# Schmutz (2012). This source is useful as it provides another perspective on NoSQL databases, including a background into NoSQL and why you might use it instead of traditional RDBMS, it describes the different NoSQL database types and compares them to the RDBMS. It also describes a different way of programming called Polyglot Programming which is used to take advantage of many different languages advantages when tackling different problems and how this can be applied to database systems. This is useful as it shows different stratagems with the task of most efficiently resolving the big data issue. Therefore I will use this when researching database systems that handle big data and in my report when evaluating the artefact.

# Wayner (2011). This source details the developments taking place in terms of NoSQL databases with Oracle releasing its version which provides many of the functions that pioneers have been experimenting with, while keeping a professional solid database finish and the performance needed to handle big data. It also describes testing of this new database, when compared to other NoSQL databases it fared very well and shows how ACID (Atomicity, Consistency, Isolation, Durability) is clearly very useful when trying to get the most performance and reliability out of a database. Therefore I will use this when deciding which database technologies to use in my artefact and in the report when evaluating the artefact.

# CUBRID (2011). This goes into great detail on the fundamentals of NoSQL as well as comparing this to relational database management systems (RDBMS) as well as a list of the popular NoSQL storage systems. This will be useful should I choose to use NoSQL when developing my artefact as a greater understanding of the underlying systems will greatly improve the successfulness of the finished system. Therefore I will use this when deciding which of the database technologies I have researched to include in my artefact, this will also be included in the report.

# LINUX For You (2012). This source details on the alternatives to traditional OLTP systems when dealing with big data, such as NoSQL and goes on to show the benefits and drawbacks of using such a system. It also makes it clear why such systems are needed now more than ever and introduces NewSQL a collective name for alternative database systems, it goes on to illustrate the features and the uses for such systems. This will be useful when choosing what systems I plan to use for my artefact and will help me make a better decision and therefore a better artefact. Therefore this will be used when researching database systems to include in my artefact and it will be useful when evaluating the artefact.

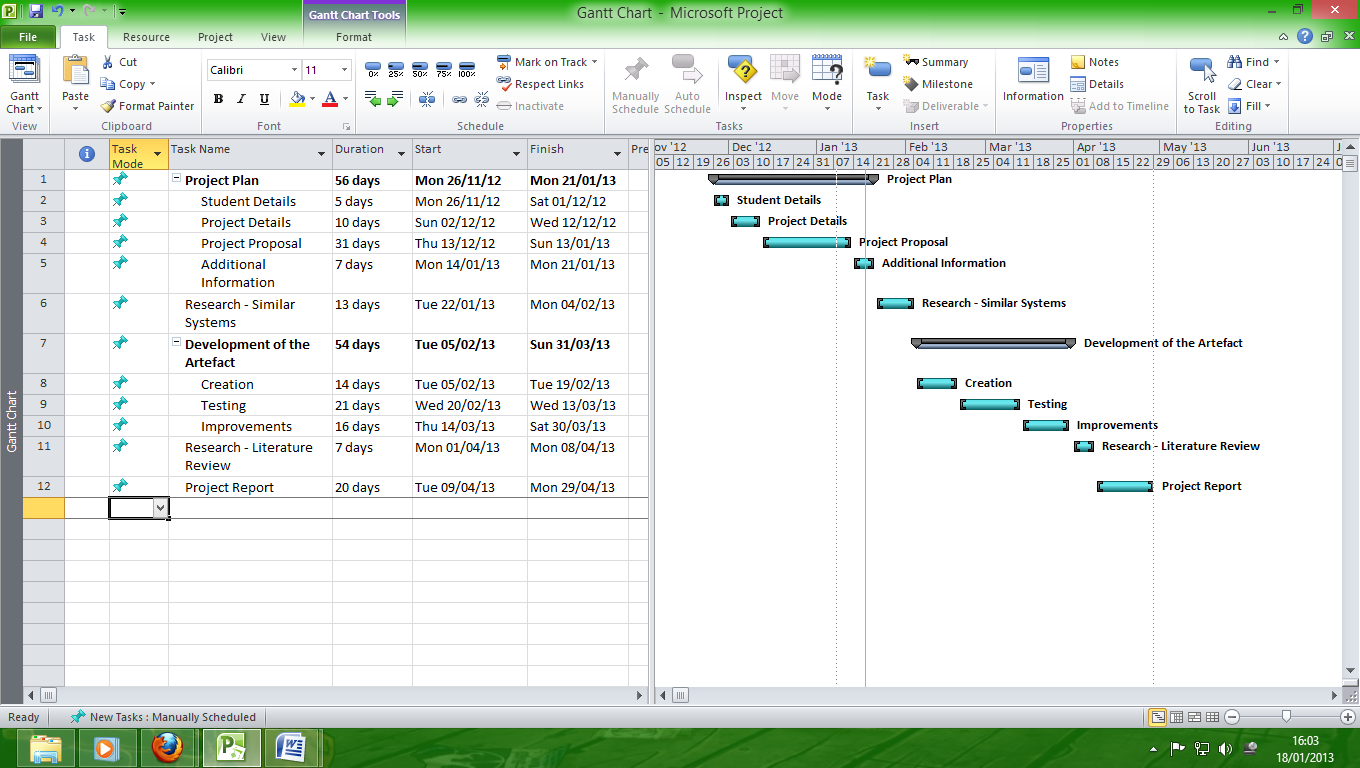
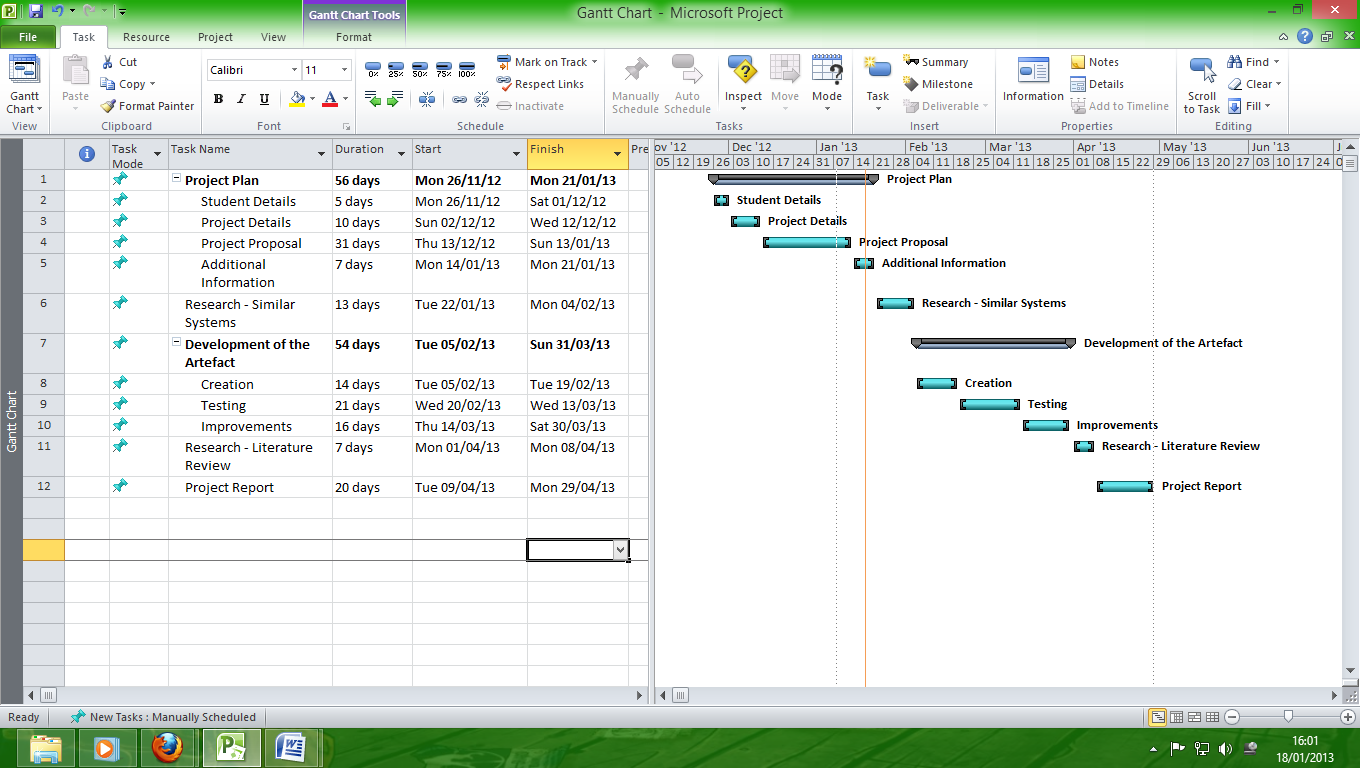
# TPC (2012). This source details the benchmarks for a wide variety of hardware and software configurations which are purposefully made to provide the best performance in online transaction processing (OLTP). This will be invaluable to me when testing my artefact in order to get context from the existing systems businesses use today, I will compare my artefacts testing results to these and evaluate it with this information in mind. Therefore I will use this when testing my artefact in order to compare it against other big data systems, it will also be included in the report when evaluating the artefact.

# 451 Research (2011). This source briefly shows the current market developments for database systems and gives an insight into the database systems emerging in order to handle the issue with the ever higher requirements needed from today's database systems. This will be useful when deciding what systems I intend to use in order to create my artefact and in doing so will help me build the most efficient system possible. Therefore I will use this when researching database systems and will be included in the report.

**Artefact**

I will create a design that will be tested to ascertain whether or not it can adequately handle incredibly large size of data. In doing so I will show which of the emerging database technologies I have used in order to create a system that can handle incredibly large size of data. The results I gain from testing such a system will allow me to illustrate whether these emerging database technologies allow for building of a system to handle big data in a realistic way, if it is practical and advantageous for a business to invest in the creation of their own ‘big data machine’.

**Plan/Schedule**



Project Plan:-

Here I will plan out my project it will include my topic area and the title of my project, my academic question, aims and objectives and my proposed artefact. It will also give background information to my project, why I chose the project, what I intend to do and go into more detail with my academic question. There will be a literature survey which will show a brief account of the available material on my chosen topic which will be fully referenced in Harvard style. I will then justify my artefact and show how it relates to my academic question, it will also show detailed scheduling in the form of a Gantt chart illustrating major milestones and deadlines. It will include an ethical considerations form signed by my supervisor, a detailed account of any ‘special’ hardware/software needed as well as details on my client.

Research, Similar Systems:-

Here I will put together a comprehensive bibliography of the emerging database systems that I will be looking to use in creation of my artefact. This will help me to decide which systems work best together and what is feasible in the time frame and other limitations of this project.

Development of the Artefact:-

Here I will take what I have learned based on the research done beforehand and create my artefact from chosen systems that are designed to handle incredibly large size of data in a way that I believe will yield the best results. I will then test this design by creating test data and using targeted queries to see if it will be practical and beneficial for businesses to consider investing in a ‘big data machine’. I will then continue to improve the design during the development process in order to create a more efficient artefact.

Research, Literature Review:-

Here I will illustrate all of the material found for my project, these sources will be reviewed at length as well as compared, contrasted and evaluated in order to show my understanding of these sources and to ascertain the quality of the research materials.

Project Report:-

Here I will include details of the completed artefact, this brief documentation will summarise exactly what systems were used in construction of the artefact, an illustration of how this design handles incredibly large size of data as well as a review of the testing performed. I will discuss based on the evidence, my project; how it relates to the academic question and based on this will conclude with an answer to my academic question. Then I will evaluate my artefact, the findings gained from testing the artefact and the process which culminated in the creation of the artefact. I will then give my conclusions of the project drawing on the initial aims of the project as well as the academic question. I will include a bibliography of all the research materials in Harvard style as well as any appendices to be included to support my Project Report.

**Bibliography**

# Arneil, J. (2012) *My Year With Exadata* [online]. e-DBA. Available at:

# <http://2012.ukoug.org/presdisplayfile.asp?prs\_prsid=7038&filename=UKOUG%5FOCT%5FMy%5Fyear%5Fwith%5FExadata%2Eppt>.

# Carr, J. (2012) *Big data’s little secret: Hadoop isn’t the end-all-be-all. VentureBeat.* [online]. 18 December 2012 [Accessed 04 January 2013]. Available at: <http://venturebeat.com/2012/12/18/hadoop/>.

# Clegg, D. (2012) *Formulating Your Big Data Strategy: It's Not About Technology - It's About Business Value; It's Not About Data - It's About Insight from Data* [online]. Acunu. Available at: <http://2012.ukoug.org/presdisplayfile.asp?prs\_prsid=7013&filename=121203UKOUG%2Epdf>.

# Howard, P., Sowerby, D. (2012) *The Next BIG Thing?* [online]. SolStonePlus. Available at: <http://2012.ukoug.org/presdisplayfile.asp?prs\_prsid=7491&filename=SolStonePlus%5F%5FUKOUG%5FConference%5FEndeca%5FPresentation%5F2012%2Epptx>.

Kutrovsky, C. (2012) *Exadata Data warehousing: Leveraging Parallel Query* [online]. Pythian. Available at:

<http://2012.ukoug.org/presdisplayfile.asp?prs\_prsid=7482&filename=Christo%5FKutrovsky%5F%2D%5FExadata%5FDatawarehousing%5FLeveraging%5FParallel%5FQuery%5F%2D%5F2012%2Epptx>.

Logan, P., Abbey, M. (2012) *Applying Traditional DBA Skills to Oracle Exadata* [online]. Pythian. Available at:

<http://2012.ukoug.org/presdisplayfile.asp?prs\_prsid=7213&filename=exadata%5Fukoug12%5Fpost%5Fabbey%5F%2D%5FPaul%5FLogan%2Epptx>.

Powell, R. (2011) *Incorporating Big Data into an Enterprise Information Architecture. BeyeNetwork* [online]. 24 October 2011 [Accessed 12 December 2012]. Available at:

<http://www.b-eye-network.com/view/15480>.

Schmutz, G. (2012) *NoSQL Databases for Implementing Data Services – Should I Care?* [online]. Trivadis. Available at:

<http://2012.ukoug.org/presdisplayfile.asp?prs\_prsid=7424&filename=nosql%2Ddatabase%2Dfor%2Dimplementing%2Ddata%2Dservices%2Dshould%2DI%2Dcare%2Epdf>.

# Wayner, P. (2011) *First look: Oracle NoSQL Database. InfoWorld.* [online]. 16 November 2011 [Accessed 06 December 2012]. Available at: <http://www.infoworld.com/d/data-explosion/first-look-oracle-nosql-database-179107>.

# CUBRID (2011) [*What is NoSQL for?*](http://www.cubrid.org/blog/dev-platform/what-is-nosql-for/)[online] Korea: NHN Search Solutions [Accessed 30 December 2012]. Available at: <http://www.cubrid.org/blog/dev-platform/what-is-nosql-for/>.

# LINUX For You (2012) *NewSQL — The New Way to Handle Big Data* [online] India: Electronics For You (EFY) [Accessed 02 January 2013]. Available at: <http://www.linuxforu.com/2012/01/newsql-handle-big-data/>.

Transaction Processing Performance Council (TPC) (2001) *Transaction Processing - OLTP* [online] San Francisco: TPC. [Accessed 27 December 2012]. Available at: <http://www.tpc.org/tpcc/results/tpcc\_perf\_results.asp>.

# 451 Research (2011) NoSQL, *NewSQL and Beyond: The drivers and use cases for database alternatives* [online] New York: 451 Group [Accessed 21 December 2012]. Available at: <https://451research.com/report-long?icid=1651>.

**Additional Information**

**UNIVERSITY OF WOLVERHAMPTON**

**SCHOOL OF TECHNOLOGY**

**ETHICAL CONSIDERATION FOR STUDENTS STUDYING TAUGHT PROGRAMMES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section 1: Your details** | | | |
| **First Name & Surname:** | James Braznell | **Student No:** | 1007022 |
| **Course:** | Computer Science | | |
| **Project Title** | What database technologies are emerging to handle incredibly large size of data? | | |
| **Supervisor:** | Sherin Nassa | | |

|  |  |
| --- | --- |
| **Section 2: Your Project Topic** | |
| **2.1** What problem is this project addressing? (100 words or less) | My project hopes to solve the issue of big data, big data is information taken from a variety of sources such as social media feeds, sensor data from products and web blogs etc. This data can be used to gain very useful information about how products are being perceived as well as potential customers thoughts of your business actions. As useful as this data is, it is generally unstructured and can add up to many terabytes of data to be stored. Therefore ways of making sense of and efficiently exploiting such data would be very beneficial to a company. |
| **2.2** Will information or artefact resulting from your project be available externally to the University? | Yes/No?  No |
| **2.2.1**  **If you answered ‘yes’ to 2.2,**  Will any such information place anyone at risk or possibly result in any action that might be detrimental to their wellbeing? (See guidelines) | Yes/No? |
| **2.2.2**  In what format will the information or artefact be made available? | The information will be stored in data storage system and targeted queries will bring back the relevant data. An analytics program can do this graphically to provide better sense of the data collected. |

|  |  |
| --- | --- |
| **Section 3: Method of Data Collection** | |
| **Please attach samples with this form if you intend to do interviews, surveys, or questionnaires.** | |
| **3.1** Does any part of your proposed project involve human participants?  If No go to Section 4. | Yes/No?  No |
| **3.1.1**  **If you answered ‘yes’ to 3.1,**  Is the sole involvement of human participants in order to provide opinions to support the specification or testing of an artefact to be produced as an outcome of the project? | Yes/No? |
| **3.1.2**  **If you answered ‘yes’ to 3.1.1,**  Does this artefact/information have any characteristics which might be detrimental to the wellbeing of any human participants in your project? If so, explain. | Yes/No? |
| **3.2**  **If you answered ‘yes’ to 3.1,**  Are there other ways you might meet your project aims without involving human participants? If not, why?  If yes discuss with your Supervisor how you will achieve this and go to section 4. |  |
| **3.2.1**  How will you select your participants? |  |
| **3.2.2**  How many participants will you contact? |  |
| **3.2.3**  How will you approach potential participants? E.g. email, letter, face to face? |  |
| **3.2.4**  Are your participants adults? (over 18 and competent to give consent) If no, answer 3.2.5 | Yes/No? |
| **3.2.5**  Are your participants children or adults over 18 and not competent to give consent? If yes, why is it necessary to involve these participants? (See guidelines)  Explain how you will ensure parental/guardian consent. |  |
| **3.2.6**  Are you offering any incentives to any of your participants, financial or otherwise? (See guidelines) | Yes/No? |
| **3.2.7**  How much time do you estimate will be needed from any participants? (See guidelines) |  |
| **3.2.8**  Please list the method of data collection and analysis intended to be used |  |
| **3.2.9**  Will all of the data collected contribute towards your results? |  |

|  |  |
| --- | --- |
| **Section 4: Confidentiality and data handling** | |
| **Please read methods of ensuring confidentiality in the guidelines.** | |
| **4.1** Will you ensure the anonymity of data collected from/and about participants? | Yes/No  Yes |
| **4.2** Will you store/protect data collected from individuals e.g. password protected files? | Yes/No  Yes |
| **4.3** Once your project is complete and information is no longer needed, will you destroy your data? | Yes/No  Yes |
| **4.4** Will anyone else have access to the data collected? | Yes/No  No |
| **If so,**  (i) please name the individuals and/or groups that will have access;  (ii) why is access being given to those listed in (i)? |  |

|  |  |
| --- | --- |
| **Section 5: Working with other parties and companies** | |
| **5.1** Will you be using data on subjects held by another party or organisation? | Yes/No  No |
| **If Yes,**  (i) Please give details.  (ii) How will you gain access to this information? |  |
| **5.2** Do you require written permission from a company, organisation or location, e.g. an employer or local authority? | Yes/No  No |
| **If Yes,**  (i) Please complete an [external agreement form](file://C:\Documents%20and%20Settings\in5541\Local%20Settings\Temporary%20Internet%20Files\in7475\Local%20Settings\Documents%20and%20Settings\in7475\Local%20Settings\Documents%20and%20Settings\in5505\Local%20Settings\Temporary%20Internet%20Files\Content.Outlook\0JYTHJ5A\External%20Computing%20Project%20Agreement%20Form.doc) and include this with your submission. |  |
| **NB: If working with another organisation or company please familiarise yourself with their Health & Safety procedures.** | |

**Things you must be aware of:**

**Data Protection Act**: <http://www.ico.gov.uk/what_we_cover/data_protection.aspx>

**Freedom of Information Act**: <http://www.opsi.gov.uk/Acts/acts2000/ukpga_20000036_en_1>

[University of Wolverhampton Ethical Approval Procedural Guidelines](http://www.wlv.ac.uk/PDF/aca-pols-ethics-scrutiny.pdf)

**Checklist:**

1. If you are using a questionnaire or interview sheet please include a list of sample questions with your submission.

2. In addition, please include an introductory cover letter stating some information about you, your project proposal and how your data will be used.

3. If you are undertaking a project involving a company or organisation you will need to show that you have approval from that organisation. Please include a completed copy of the [External Agreement Form](file://C:\Documents%20and%20Settings\in5541\Local%20Settings\Temporary%20Internet%20Files\in7475\Local%20Settings\Documents%20and%20Settings\in7475\Local%20Settings\Documents%20and%20Settings\in5505\Local%20Settings\Temporary%20Internet%20Files\Content.Outlook\0JYTHJ5A\External%20Computing%20Project%20Agreement%20Form.doc).

|  |  |
| --- | --- |
| **Student’s Declaration**  Sign and date against **one** declaration **only** | |
| **Category 0.** I have answered  ‘No’ to questions **2.2** and **3.1**.  My project involves no human participation except for myself and I agree to ensure that any information or artefact produced will not be available outside the University. | James Braznell  21 January 2013 |
| **Category A1.** I have answered (delete one from each block)   |  | | --- | | ‘Yes’ to questions **2.2**, **4.1**, **4.2** and **4.3** and ‘No’ to questions **2.2.1** and **4.4**. | | ‘No’ to question **2.2** |   **and**   |  | | --- | | ‘Yes’ to questions **3.1**, **3.1.1** and **3.2.4** and ‘No’ to question **3.1.2** | | ‘No’ to question **3.1** |   My project involves limited human participation and I agree to ensure that   1. any such participation is not detrimental in any way to the interests of the participants; 2. all information collected as a part of the project will be handled in accordance with the answers that I gave to question **4**; 3. No information or artefacts which may place anyone at risk or be detrimental to their wellbeing will be made available outside the University. |  |
| **Category A2.** I have answered  ‘Yes’ to question **3.1**  **or** I have answered  ‘Yes’ to question **2.2**  and my answers to subsequent questions prevent the project being classified as A1.  My project involves human participation and may present some risk to participants. I have considered alternative means of pursuing the project which do not entail this risk but believe that there is no practicable alternative. I agree to ensure that I take all necessary steps to minimise risks to participants and third parties. I agree not to proceed with any activities involving human participation until I have received approval from the Department Ethics Panel. |  |
| **Category B-E.** My project does not conform to Category 0, A1 or A2. I have considered alternative means of pursuing the project which do not entail risk to human participants but believe that there is no practicable alternative to the proposal made. I agree to ensure that I take all necessary steps to minimise risks to participants. I agree not to proceed with any activities involving human participation until I have received approval from the School or University Ethics Committee, as appropriate. |  |

|  |  |
| --- | --- |
| **Supervisor’s Declaration**  Sign and date against **one** declaration **only** | |
| **Category 0 or A1.** I concur with the classification of this project as **0** or **A1** and authorise continuation of the project. I have forwarded a copy of this form to the Department Ethics Panel for monitoring purposes. |  |
| **Other.** I believe that this project should be classified other than **0** or **A1.** I do **not** authorise continuation of the project until approval has been received from the appropriate Ethics Panel or Committee. I have forwarded a copy of this form to the Department Ethics Panel for consideration. |  |

**FOR SUPERVISOR/PANEL/COMMITTEE USE ONLY:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFICATION ALLOCATED BY SUPERVISOR** | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **0, A1** |  | **Supervisor Action:** Authorise and forward to DEP | | | | | | | | | | | | | | | | | | Date | | | |  | | | | |
| **DEP Action:** File for possible monitoring | | | | | | | | | | | | | | | | | | | Date | | | |  | | | | |
|  | | Selected for monitoring | | | |  | (tick) | | | | | | | | | | | | Date | | | |  | | | | |
| Classification agreed? | | | | Yes |  | No |  | | | If ‘No’, give: | | | | | | | | | | | | | | | |
| reason | | | |  | | | | | | | | | | | | | | | | | | | | | |
| action | | | |  | | | | | | | | | | | | | | | | | | | | | |
| **Other** |  | **Supervisor Action:** Refer to DEP for decision | | | | | | | | | | | | | | | | | | Date | | | |  | | | | |
| **CLASSIFICATION ALLOCATED BY DEPARTMENT ETHICS PANEL** | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **0, A1** | |  | **DEP Action:** Project authorised to continue | | | | | | | | | | | | | | | | | Date | | | |  | | | |
| **A2** | |  | **Considered by DEP below** | | | | | | | | | | | | | | | | | Date | | | |  | | | |
| **2.2** Is any risk associated with access to project acceptable in context? If no, give reasons below: | | | | | | | | | | | | | | | | | Yes | | |  | | | No | |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | | |
| **3.1** Is involvement of human participants justified? If no, give reasons below: | | | | | | | | | | | | | | | | | Yes | | |  | | | No | |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | | |
| **3.3** Is experimental method acceptable with regard to risk and inconvenience to participants? If no, give reasons below: | | | | | | | | | | | | | | | | | Yes | | |  | | | No | |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | | |
| **4** Are arrangements for confidentiality and data protection appropriate? If no, give reasons below | | | | | | | | | | | | | | | | | Yes | | |  | | | No | |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | | |
| **5** Do arrangements for working with external bodies protect interests of participants and the external bodies? If no, give reasons below | | | | | | | | | | | | | | | | | Yes | | |  | | | No | |  |
|  | | | | | | | | | | | | | | | | | | | | | | | | | |
| **DEP Action:** Continuation of project approved: | | | | | | | | Yes | | |  | | No | |  | | Date | | | |  | | | | |
| Conditions: | | | | | | | | | | | | | | | | | | | | | | | | | |
| **Other** | |  | **DEP Action:** Refer to School Ethics Committee | | | | | | | | | | | | | | | | | Date | | | |  | | | |
| **CLASSIFICATION ALLOCATED BY SCHOOL ETHICS COMMITTEE** | | | | | | | | | | | | | | | | | | | | | | | | | |
| **0, A1** | |  | **SEC Action:** Continuation of project approved | | | | | | | | | | | | | | Date | | | |  | | | | |
| **A2, B** | |  | **Considered by SEC below** | | | | | | | | | | | | | | Date | | | |  | | | | |
| **2.2** Is any risk associated with access to project acceptable in context? If no, give reasons below: | | | | | | | | | | | | | | | Yes | | |  | | | | No |  |
|  | | | | | | | | | | | | | | | | | | | | | | | |
| **3.1** Is involvement of human participants justified? If no, give reasons below: | | | | | | | | | | | | | | | Yes | | |  | | | | No |  |
|  | | | | | | | | | | | | | | | | | | | | | | | |
| **3.3** Is experimental method acceptable with regard to risk and inconvenience to participants? If no, give reasons below: | | | | | | | | | | | | | | | Yes | | |  | | | | No |  |
|  | | | | | | | | | | | | | | | | | | | | | | | |
| **4** Are arrangements for confidentiality and data protection appropriate? If no, give reasons below | | | | | | | | | | | | | | | Yes | | |  | | | | No |  |
|  | | | | | | | | | | | | | | | | | | | | | | | |
| **5** Do arrangements for working with external bodies protect interests of participants and the external bodies? If no, give reasons below | | | | | | | | | | | | | | | Yes | | |  | | | | No |  |
|  | | | | | | | | | | | | | | | | | | | | | | | |
| **SEC Action:** Continuation of project approved: | | | | | | | Yes | | |  | | No | |  | | | Date | |  | | | | |
| Conditions: | | | | | | | | | | | | | | | | | | | | | | | |
| **Other** | |  | **SEC Action:** Refer to University Ethics Committee | | | | | | | | | | | | | | | | Date | |  | | | | |

**Guidelines**

**Section 1: Categorisation for ethical approval**

**Category 0:** There are no third parties directly involved in the project and any artefacts produced by the project will not be accessible to a general audience.

**Category A1**

Projects involving human volunteers are involved solely for the purposes of:

- providing data to inform the specification of an artefact

- testing the usability or fitness for purpose of an artefact

where the nature of that artefact or its use will present no risk to the volunteers

and, if any artefact is accessible to a general audience, access to that artefact will present no risk.

**Category A2**

Projects involving human volunteers other than those defined in category A1 but not in activities defined in other categories or if any artefact is accessible to a general audience, access to that artefact may present some risk.

**Category B**

Projects involving human volunteers including potential risk, for instance,

studies using new research methodologies, studies involving certain vulnerable

populations or therapeutic interventions or other significant risk to anyone involved in

the research (but not including trials of artefacts intended for therapeutic purposes).

**Category C**

Research being conducted by staff or postgraduate research students involving

Patients, clients staff, records etc. within the sphere of the NHS, Social Services, etc (but not including clinical trials of medicinal or related products).

**Category D**

Research being conducted by undergraduate or taught postgraduate students involving

Patients, clients staff, records etc. within the sphere of the NHS, Social Services, etc (but not including clinical trials of medicinal or related products).

**Category E**

Clinical trials of medicinal or related products involving patients or healthy volunteers as direct users of the product.

**Question 2.2:** You should answer yes if your artefact, product or information might be of direct risk or might lead or encourage people to alter their behaviour in a way which would be detrimental to them. Examples of direct potential risk might be a machine that could injure someone if it malfunctioned or a web resource which contained information which if it was misused would lead to risk (for instance, children’s identities or addresses). Examples of artefacts which might encourage detrimental behaviour could be a web resource offering alternatives to expert (such as GP or lawyer) advice or products which purport to have a therapeutic effect.

**Question 3.2.5:** As a general principle, all participants should be informed of their role in the experiment and freely consent (in writing) to it, which implies competence to give consent. Very occasionally it may be necessary to undertake an experiment without consent, or with participants who are not competent but then any decision about the acceptability of the proposal would be taken on the basis of the absolute benefit of the experiment in a wider context, and it would have to be established that there was no alternative.

**Question 3.2.6:** With regard to freedom of consent, it likely that this principle would be breached of the participants were subject to some kind of inducement or coercion, however minor. For instance, it is likely that participants who were under the management of the person undertaking the experiment would be considered to be under a degree of coercion.

**Question 3.2.7:** It may be considered that expecting a participant to spend undue time or effort participating in an experiment would be detrimental to the interests of that person, particularly where the results of the work offered no clear benefits. It may be appropriate to compensate participants for their time, but it is not acceptable to offer inducements to participate.

**Section 4 Anonymity:**

It is to be expected that due care and attention be paid to protecting information about individuals. Depending on the nature of the experiment, the following may be considered.

* Type 1: Complete anonymity of participants (i.e., You will not meet, or know the identity of participants, as they are part of a random sample and are required to return responses with no form of personal identification)?
* Type 2: Anonymised samples or data (i.e., an irreversible process whereby identifiers are removed from data and replaced by a code, with no record retained of how the code relates to the identifiers. It is then impossible to identify the individual to whom the sample of information relates)?
* Type 3: De-identified samples or data (i.e., a *reversible* process whereby identifiers are replaced by a code, to which you retain the key, in a secure location)?
* Type 4: Subjects being referred to by pseudonym in any publication arising from the project?
* Type 5: Any other method of protecting the privacy of participants? (eg. use of direct quotes with specific, written permission only; use of real name with specific, written permission only)

For this project I will be using several different technologies to handle big data, the specific technologies that will be used in my artefact will be selected based on my research into those technologies. Based on my current research the technologies available to me are Hadoop file system, NoSQL databases such as Oracles version, NewSQL databases, MapReduce and analytics programs such as Acunu Analytics.

My supervisor Sherin Nassa will be my client