**Project Interim 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: 1st March 2013

Project Title:

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

Contents:

* Title Page, Student Details, Project Title (page 1)
* Academic Question, Aim(s), Objectives and the Artefact (page 2-3)
* Initial Literature Review (page 4-8)
* Project Progress (page 9)
* Project Plan (page 9)

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.

**Literature Review**

Borkar, R, V., Carey, J, M. and Li, C. (2012). Big data platforms: What's next?. *XRDS: Crossroads*, [online]. **19**(1), pp.44-49 [Accessed 15 February 2013]. Available at:

<http://delivery.acm.org.ezproxy.wlv.ac.uk/10.1145/2340000/2331057/p44-borkar.pdf?ip= 134.220.1.139&acc=ACTIVE%20SERVICE&CFID=281029974&CFTOKEN=14962939&\_\_acm\_\_=1361652571\_a266ed3632371081205a63df4ca9716f>

# This source describes the current status of big data, how it is being used, the history behind it, the source describes several technologies that have emerged to deal with big data such as MapReduce framework, Hadoop, BigTable, dynamo, HBase, Cassandra, Hive, Pig. The source then looks the future and introduces its own system ASTERIX, a big data management system (BDMS), the system is described and is compared to several other big data technologies showing how big data systems have advanced over the years. The source is quite useful as it shows what big data is used for, the technologies that deal with it and a new technology that takes advantage of the developments in big data technology, similar to the work of Tene and Polonetsky (2012), both sources describe big data in detail however this source focuses more on the technology for handling big data while the previously referenced source focuses more on the issues with big data. This source will help me answer my academic question and achieve my aims when it is used to help create my artefact and to make a better decision on its configuration as well as in the final report for when I evaluate the completed artefact.

Bhatewara, A., Waghmare, K. (2012). Improving Network Scalability Using NoSql Database. *International Journal of Advanced Computer Research*, [online]. **2**(6), pp.488-490 [Accessed 21 February 2013] Available at: <http://www.theaccents.org/ijacr/papers/conference/icett2012/87.pdf>

# The source describes technologies more suited to the unstructured nature of big data in this case for use in the cloud such as NoSQL database and Cassandra, it shows the advantages and disadvantages of each and how they compare to a RDBMS. This source is useful as it shows that if higher scalability is needed without upgrading hardware, these technologies are a viable option for this, similar to the work of Stonebraker (2012), they both show technologies that can give the user greater performance when dealing with big data however this source evaluates NoSQL and Cassandra while comparing them to RDBMS which the previously referenced source does not. This source will help me answer my academic question and achieve my aims by being used as reference when creating my artefact and when evaluating it in the final report.

Davis, M. A. (2012). NoSQL Equals NoSecurity. *InformationWeek*, [online]. Apr 9, 2012, pp.29-31 [Accessed 26 February 2013] Available at:

<http://search.proquest.com.ezproxy.wlv.ac.uk/docview/1011104685/fulltextPDF?accountid=14685>

The source shows that while technologies like NoSQL and Hadoop are more widely used for their ability to handle big data, it is clear that security is an issue that may have been overlooked. It details how several forms of NoSQL databases are being adopted as well as the use of MapReduce and BigTable to handle big data however security is very lax due to the advancements in such technologies coming from the developers who have so far have not seen fit to add proper security measures to such technologies. It recommends security controls for developers to implement for the database as well as security recommendations for the operating system (OS) and concludes that companies must take more responsibility in terms of security. This source is very useful as it provides a great deal of information on the security issues of big data technologies especially NoSQL databases which is similarly detailed in the work of Tene and Polonetsky (2012). However in contrast to this source the previously referenced source details a background on big data as well as the benefits of big data analysis which this source does not. Therefore this source is very useful as a change from the benefits of big data technologies and I will help me answer my academic question and achieve my aims by using during the creation of my artefact and in the report when I consider the security issues with big data and my artefact.

Kutrovsky, C. (2012) Exadata Data warehousing: Leveraging Parallel Query. *Cover Art: UK Oracle User Group Conference 2012* [online]. ICC Birmingham 3-5 December. Pythian.

Password: CONF2012OUG

[Accessed 08 December 2012] 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>.

The source details the inner workings of Exadata, specifically how the data moves in order to get a result, it describes parallel query concepts including a big sort, group by and join which is useful when dealing with very large tables. There are many useful points in this source that helps to shed light on how Exadata works to give the best result, compared to other sources this source gives the reader a fair amount of information on Exadata similar to the work of Logan and Abbey (2012), which also provides an adequate amount of information on big data and oracles Exadata. This source will help me answer my academic question and achieve my aims by being used in the final report when comparing my artefact to Exadata.

Logan, P., Abbey, M. (2012) Applying Traditional DBA Skills to Oracle Exadata. *Cover Art: UK Oracle User Group Conference 2012* [online]. ICC Birmingham 3-5 December. Pythian.

Password: CONF2012OUG

[Accessed 06 December 2012] Available at:

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

The source compares and contrasts a relational database management system (RDBMS) to Exadata and how Exadata works to get the most performance possible, backup considerations, the internals of the Exadata machine, functions for compression, performance and storage, this shows what is new with Exadata and how it can help deal with big data. There are useful points in the source which when compared to the work of Howard and Sowerby (2012), which in contrast to this source provides the reader with less information than is expected. This source will help me answer my academic question and achieve my aims by helping tocomparing my artefact against Exadata in the final report.

Mone, G. (2013). Beyond Hadoop. *Communications of the ACM*, [online]. **56**(1), pp.22-24 [Accessed 12 February 2013] Available at:

<http://wk6kg9sd8m.search.serialssolutions.com/?ctx\_ver=Z39.88-2004&ctx\_enc=info%3A ofi%2Fenc%3AUTF-8&rfr\_id=info:sid/summon.serialssolutions.com&rft\_val\_fmt=info:ofi/fm t:kev:mtx:journal&rft.genre=article&rft.atitle=Beyond+Hadoop.%28Apache+Hadoop%27s+continuous+evolution%29%28News+%2F+Technology%29&rft.jtitle=Communications+of+the+ACM&rft.au=Mone%2C+Gregory&rft.date=2013-01-01&rft.pub=Association+for+Computin g+Machinery%2C+Inc&rft.issn=0001-0782&rft.eissn=1557-7317&rft.volume=56&rft.issue=1 &rft.spage=22&rft.externalDBID=n%2Fa&rft.externalDocID=316952108>

# This source shows how Hadoop continues to be used as a leading technology for handling big data, it describes the history and structure of Hadoop as well as mentioning several technologies that add onto or focus more on certain aspects of Hadoop such as, MapR, Pregel, GraphLab, HBase, Cloudant, Dremel for example and gives a description of each. This source is useful as it shows a different side to Hadoop where although it will handle the issue of big data but it alone does not allow for the tools needed to reap the benefits from analysing such data, similar to the work of Borkar et al. (2012), both sources give detailed information on several technologies that handle big data however this source does not go into detail on big data itself while the previously referenced source does and this source focuses on hadoop and its offshoots, while the previously referenced source branches out to several different big data technologies. This source will help me answer my academic question and achieve my aims when I use it as guidance when creating my artefact and in the report when I evaluate the artefact.

Schmutz, G. (2012) NoSQL Databases for Implementing Data Services – Should I Care? *Cover Art: UK Oracle User Group Conference 2012* [online]. ICC Birmingham 3-5 December. Trivadis.

Password: CONF2012OUG

[Accessed 10 December 2012] Available at:

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

The source gives a perspective on NoSQL databases, a background, reasons for using it instead of a RDBMS, a description of different NoSQL databases, a comparison to RDBMS, it shows a different form of programming called Polyglot Programming used for taking advantage of many different languages advantages when tackling different problems and how this can be applied to database systems. This source is similar to the work of Bhatewara and Waghmare (2012). Both sources show in great detail technologies that handle the issue of big data and details why these are used over a traditional RDBMS, this source will help me answer my academic question and achieve my aims when constructing my artefact as well as in the final report.

Stonebraker, M. (2010). SQL databases v. NoSQL databases. *Communications of the ACM*, [online] **53**(4), pp.10-11 [Accessed 22 February 2013] Available at:

<http://mags.acm.org.ezproxy.wlv.ac.uk/communications/201004/?folio=10&CFID=281029974&CFTOKEN=14962939#pg1>

# The source describes the reasons for using a NoSQL database and how to get better performance for online transaction processing (OLTP) databases by removing the 'overhead' or traditional aspects of Atomicity, Consistency, Isolation, Durability (ACID) which is grouped into four parts in OLTP databases, for example; logging, locking, latching and buffer management these points are given descriptions for clarity. Therefore this source is quite useful as it shows the differences between RDBMS and NoSQL databases and how NoSQL allows for greater performance, similar to the work of Schmutz (2012) which also gives its perspective on NoSQL databases and similarly compares it to RDBMS, however this source shows how NoSQL database gives the user greater performance over RDBMS which the previously referenced source does not. This source will help me answer my academic question and achieve my aims by using this source when creating my artefact and in the final report when I evaluate my artefact.

Tene, O., Polonetsky, J. (2012). Privacy in the Age of Big Data: A Time for Big Decisions. *Stanford Law Review Online*, [online] **64**(63), pp.63-69 [Accessed 19 February 2013].

Available at: <http://www.stanfordlawreview.org/online/privacy-paradox/big-data>

The source details a background on big data and how it has relatively recently come into the spotlight, it show the benefits big data can grant such as in the health sector, data can be analysed to attribute health issues to previously unknown causes, big data can also be analysed to predict flu outbreaks or energy demands for example. The concerns with big data are detailed such as protecting privacy and data security and therefore governments and businesses must be aware of this issue. The source therefore is quite useful and raises some issues that come with big data as well as its benefits, similar to the work of Borkar et al. (2012). This source gives the reader a background on big data, benefits etc which is similar to the previously referenced source, however this source also shows the privacy and security concerns with big data which is in contrast to the previously referenced source as it does not mention this.This source will help me answer my academic question and achieve my aims by providing me with background knowledge for the project but also for use in the final report.

**Project progress**

At this point in the project i have completed the project plan, including the student details, project details, project proposal and additional information. I have completed my initial research into similar systems or 'big data systems', i am currently in the process of developing my artefact, the main systems i will use have been installed and i have begun testing of these systems. In terms of my Aims, i am progressing towards completing my first aim, the objectives i have so far completed are:

* 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

**Project Plan**

My progress so far is in line with my Gantt chart and the only major decision made during the planning stage of the project was to not meet the expectations of the British Computer Society (BCS).