-

Working With Newer Data Management Technologies

Clare Stanier
Staffordshire University
Beaconside
Stafford ST18 0AD
c.stanier@staffs.ac.uk

Emmanuel Isitor
Staffordshire University
Beaconside
Stafford ST18 0AD
I000299a.student.staffs.ac.uk

Special thanks to Aleksej Fomin and Ezra Chatla who gave permission for their work to be discussed in this paper:

Abstract

Data management technologies are changing rapidly and this presents a significant challenge for database teaching. There is a requirement to teach traditional relational database concepts and to ensure that students are equipped with the advanced skills expected by employers. There is also a requirement to prepare students to work with newer data models and NoSQL and to understand and be able to leverage concepts such as Big Data analytics. This paper discusses the experience of working with MongoDB and MapReduce and starting to work with Hadoop in undergraduate and postgraduate teaching at Staffordshire University. It is suggested that while the amount of time that can be given to newer technologies in the undergraduate curriculum is limited, this is a subject area which has the power to capture students' imaginations and provides a good basis for undergraduate projects and Masters level dissertations.

Keywords
Big Data; NoSQL; Data Curriculum

1. Introduction

Industry data shows that although systems based on the relational data model still dominate the global database market, data management approaches and technologies designed for what is known as NoSQL and 'Big Data' are growing rapidly (Market Research Media, 2013, Gartner, 2014). Figures as to job vacancies are difficult to verify but it is clear that that this is an expanding area and one where there is currently a skills shortage (Barnes, 2014). Since the introduction of the relational data model there have been a series of 'New Waves' designed to support more efficient and/or more flexible data handling. Over time, these

newer approaches have tended either to be incorporated into relational implementations as with object-relational and XML functionality or have remained relatively peripheral to main stream data management (Java Developer 2004, Leavitt, n.d.). The Big Data era represents a different kind of challenge since the driver is not a technology looking for an application, as was arguably the case with object-oriented databases, but the data itself. Boyd and Crawford (2012) described Big Data as a socio-technical phenomenon which changes the definition of knowledge. Kambatla et al. (2014) note that data traffic grew 56-fold between 2002-2009 and that data creation and harvesting are both forecast to continue to expand rapidly. Big Data is most often defined by its characteristics (the 3 or 5 'V's (Demchenko et al. 2013) but a more useful definition here is the shorthand version given by Madden (2012) – data that is too big, too fast or too hard for existing tools to process. Data now exists in volumes and in formats that were previously unknown or inaccessible and there has been a corresponding emergence of systems and tools developed to handle the data; this provides new challenges in data handling and data analytics (Kraska, 2013; Kambatla et al. 2014) and for the data curriculum. For the purposes of this paper, we understand the term 'newer data management technologies' to include NoSQL databases and the issues involved in storing, processing and analysing Big Data.

2. The Data Curriculum

The Quality Assurance Agency (QAA) 2007 Computing Benchmark statement (undergraduate) includes databases and related topics. Of 20 BSc (Hons) Computer Science and Computing Science degrees offered by UK universities for 2014/5 (randomly selected, from both Russell Group and post-92 institutions), all but one included some version of databases/data management in their outline online curriculum, sometimes specifying relational databases. There are, however, very few undergraduate specialist awards in database/data management as compared for example with Computer Security or Computer Networking. The UCAS undergraduate course search for 2014/5 found 7 courses that had database in the title but these were HNC/HND or foundation degrees. Two undergraduate Data Science courses were found, one at the University of Warwick, one University of Bedfordshire. There were a large number of Information Systems and Business Information undergraduate honours degrees and these awards typically include a substantial element of database work. At postgraduate level there was a wide range of specialist Data Science or equivalent masters degrees. This appears to be part of a global trend with new courses in data analytics being developed in the States, in Australia, India and Eire (Patel, 2013).

This data suggests a number of conclusions: traditional database topics are a core part of the general undergraduate computing curriculum but are primarily a service subject, in the same way that all computing students study maths and programming. A number of institutions offer specialist/advanced database/data management modules but these tend to be options, often in the final year. Information Systems and related specialist degrees include more data management/database topics than other undergraduate awards. At postgraduate

level there is much more scope for specialisation. This has implications for the teaching of NoSQL/Big Data topics. Some areas such as data storage and processing may be covered as part of other specialist awards, as for example in the Staffordshire University Cloud Computing degree but as a generalisation, the database element in the undergraduate core curriculum tends to focus on established relational and object oriented database technologies. This reflects the experience at Staffordshire University where NoSQL/Big Data topics are taught as part of final year advanced database option modules and on postgraduate awards. We make the assumption that students come to these topics with a good understanding of relational and object relational design and implementation and that they have chosen database option modules because they have an interest in data management.

3. Introducing Big Data & NoSQL topics to the curriculum

3.1 Related Work

Big Data as a topic is starting to appear in subject areas as different as Quality (Antony, 2013), Health (Sherestha 2014, and Engineering (Jelinek & Bergey 2013). In computing, there have been a number of studies on introducing Big Data and NoSQL topics to the undergraduate and postgraduate curriculum. The Portland approach is to introduce Big Data concepts at an early stage, teaching MapReduce, for example, at first year undergraduate level (Grossniklaus & Maier 2012). Anderson et al. (2014) review the experience of providing a specialist data science undergraduate degree at an American university and comment on the difficulty of sustaining a separate Data Science first degree. Three core data science modules are described in detail in the paper; all deal with Data Science core concepts and in each case Big Data issues are presented to the students side by side with relational and other topics. Sattar et al. (2013) discuss the experience of an undergraduate semester long web based project which used both Oracle 11g and MongoDB. The teaching schedule shows that of the 15 week course, 10 weeks were devoted to relational concepts and only 2 to NoSQL; the remaining sessions were devoted to web interface and project work. This reflects our own experience of teaching on an advanced database systems module in which the demands of the curriculum meant that only a small part of the teaching time could be devoted to newer data management approaches and these approaches were taught side by side with traditional database topics.

4. A strategy for Introducing Big Data & NoSQL to the curriculum

The assumptions that underlie our strategy are that students come to these topics with a good grasp of relational concepts and that there is not enough curriculum space in a non specialist undergraduate award to cover newer data management approaches in depth. The focus therefore is on introducing students to relevant concepts and providing them with the underpinning knowledge to support further study. As discussed in section 5, it is our experience that Big Data/NoSQL are topics which can capture students' imaginations and

that some students would wish to explore these topics further as part of final year projects/Masters dissertations. The approach taken is outlined in Figure 1:

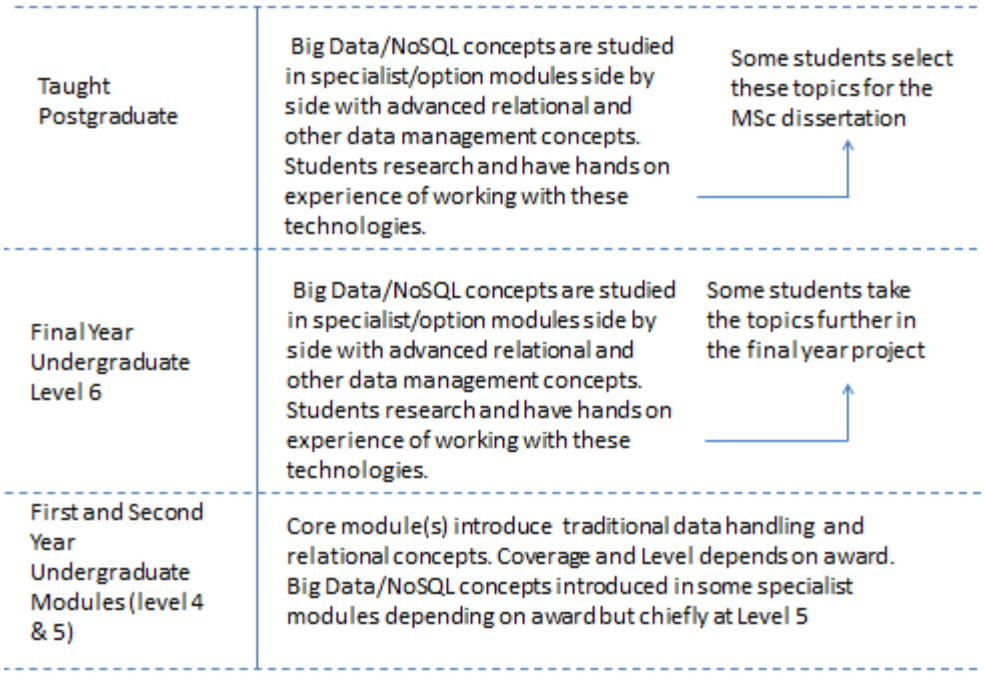

Figure 1: Coverage of Big Data and NoSQL topics

The following sections discuss our experience of working with Big Data and NoSQL.

4.1 Introduction to working with a NoSQL Datastore

Our initial experience of introducing NoSQL to the undergraduate curriculum was discussed in a paper presented to TLAD '12 (Stanier, 2012) which explained our reasons for preferring MongoDB to other open source datastores. A key factor was the ease of download and the availability of supporting documentation and tutorials. MongoDB was locally loaded on to all the machines in our specialist lab and on to a number of machines in other labs for student access outside scheduled sessions. We found that the majority of students on final year undergraduate database modules downloaded MongoDB on to their machines and no students had problems with the download. We chose to work with the JavaScript shell as we felt this would be simultaneously familiar to students but very different from relational GUIs. Our evaluation with students showed that the use of MongoDB was welcomed and that they were enthusiastic about working with a newer data management approach. Since the initial introduction of MongoDB, we have found that an increasing number of students come to the module with some prior experience of NoSQL, most usually through having worked with MongoDB on placement or having experimented with NoSQL themselves. We have not so far had any students claiming prior experience of any NoSQL database other than MongoDB but this may reflect the fact that MongoDB is now installed on most labs in the School of Computing and hence students have greater exposure to Mongo. This trend meant that in the current academic year (2013/4), we had

two groups within the undergraduate cohort taking advanced database modules. The majority of students had no experience of NoSQL and needed introductory material but there was a smaller group of students who had already advanced beyond introductory level and were interested in implementation and performance issues. We expect that in future years, as there is more coverage of NoSQL in Level 5 (second year) modules, students on final year modules will already have a basic understanding of NoSQL and it will be possible to introduce more advanced topics for the whole cohort.

When we first introduced NoSQL in 2011/2, MongoDB was initially used only in one final year undergraduate option module and the focus was on introducing students to the schema later (schemaless) element and design issues. The assessment was based on a data management case study in which some aspects of the case study lent themselves to a relational solution and some to a NoSQL solution. Students were required to develop an enterprise solution in either Oracle or SQL Server and to create and query a customer comments blog in MongoDB. For the MongoDB element, students created collections to hold documents, queried and updated documents and evaluated the differences between relational and NoSQL design. We have since extended the use of NoSQL into other modules and have started to look at a wider range of issues. As discussed in 3.1, the literature suggests that most institutions teach relational and non relational concepts side by side; we found this helpful as well understood relational concepts provided depth and context and gave students a point of reference. As an example, the syllabus for a Level 5 database security module includes the ACID protocol; this year we discussed BASE (Basically Available, Soft state Eventual consistency) and the CAP Theorem (Consistency, Availability, Partition Tolerance) in a relational and non relational context, illustrated with reference to MongoDB and developer blogs. This allowed students to relate concepts to real world data management problems. One student picked ACID/BASE for the research paper which was part of the module assessment. We have recently had a revalidation which has resulted in moving some relational topics into Level 5 modules, leaving space in final year modules for newer data management approaches.

We encountered two problems with the MongoDB installation. A minor problem was that the Data folder sometimes needed to be purged manually (worth investigating if there are unexplained problems with access). A more significant issue was that a student who wished to develop a project which focused on sharding, encountered network permission issues and had to be given extended privileges to complete the project. We are currently discussing installing MongoDB in a virtual environment to prevent permission problems in the future.

4.2 Data Sets

We have worked with large data volumes for some time as students are required to create million+ data sets for relational optimisation exercises, using either free data generators or

tutor provided data generation procedures. This is not, however, Big Data and working with data sets large enough to be described as Big Data has so far been problematic. The issue is not the availability of data sets; we have for example used the Stanford Gowalla data in a number of projects. The difficulty is load time, particularly for students working with unfamiliar functions such as mongoimport. The alternative approach is to preload data. For teaching purposes we wanted students to experience the differences between data manipulation in relational and NoSQL environments and for this reason we chose to work with small JSON datasets which could easily be created, manipulated and most importantly understood by students. As discussed in 4.3, we also propose to provide preloaded data sets for more advanced work.

4.3 MapReduce and Hadoop

Teaching MapReduce, we were limited to the MongoDB platform as we did not have an Hadoop installation for classroom use. This had some drawbacks since although MongoDB provides support for this approach, it is not optimised for MapReduce[1] and recommends the MongoDB aggregation framework instead. As we had previously decided to allow students to create their own data sets, the data used was trivial for a MapReduce operation. The advantage was that students were working with a now familiar interface and the small scale of the data made it easier for them to track what was happening. A basic tutorial was used which enabled students to map the data to Key Value pairs and then reduce the data. Comparing this with a group of students who had run the traditional MapReduce word count program, the small data set approach seemed to support a better understanding of concepts but did not give the students a sense of scale. For the next academic year, we plan to have the Hadoop ecosystem available for classroom use so that students can be introduced to MapReduce using a trivial data set and progress to working with a more realistic data set.

We have been investigating the options for Apache Hadoop both to give students a more realistic Big Data experience – we would like to introduce students to Pig and Hive and allow them to compare results with relational data manipulation – and to ensure that students who wish to use Hadoop for projects/dissertations have easy access to software. We are currently investigating 3 options. Hadoop has been installed on OS-X 10.5. We are not proposing to use this installation for classroom teaching as our data management modules use a wide range of software which will not be installed on the Macs. The OS-X version will be available for project/dissertation use. We considered a dual boot solution but this was rejected for performance and partly for security reasons. The remaining option was to work in a virtual environment. One of the authors of this paper has been working with the HortonWorks Sandbox. This is an opensource download which provides a simple interface to the Hadoop ecosystems – Hive and Pig for example are preinstalled. The system requires a virtual environment but the download is straightforward. Hortonworks

[1] Current release; this will change in future releases

provides a wide range of tutorials and support material. There are some limitations, for example on the data volumes uploaded but the author worked with data sets which although not meeting the definition of Big Data were large enough to provide a good basis for exploring the functionality provided. We are now testing the sandbox for classroom use.

5 Student Work

As discussed in section 4, NoSQL and Big Data are topics which are capable of capturing students' imaginations and which students are keen to explore further. A number of students have developed final year projects and masters' dissertations in this area, producing high quality work and we expect this to continue. To illustrate the work produced, we discuss two example projects, one undergraduate, one masters. The undergraduate project involved the creation of a performance analysis dashboard for use with MongoDB. One of the disadvantages of NoSQL databases, as compared with relational DBMS, is the relative lack of DBA tools. The project created a tool to monitor and analyse performance, integrating with 3rd party applications and using a range of programming languages and environments. A dedicated test environment was created which involved configuring multiple MongoDB servers. The Masters project was Hadoop based and investigated the causes of failure in the Name Node and the recovery process. This required the student to install the Hadoop ecosystem and develop an understanding of the Hadoop architecture as a preliminary to carrying out the project. Both projects were demanding and in both cases the installation/testing element could almost have formed a project in its own right; the challenge of working with a new and unfamiliar technology and the students' enthusiasm for the topic meant that they set themselves ambitious goals. One of our reasons for moving to use MongoDB in a virtual environment, and installing the Hadoop ecosystem on a number of machines is to support future projects.

5. Conclusion

Computing education is used to seeing topics come in and out of favour; some develop into permanent additions to the curriculum and others are relegated to history. Stonebraker (2013) described 'Big Data' as the 'Buzzword du jour' but the same article outlined the challenges that Big Data represents and the way it is changing what we can do with data. The tools that we are currently use are already evolving as more functionality is added to NoSQL databases and alternatives strategies for working with Big Data start to develop. Teaching strategies for NoSQL and Big Data will continue to be a work in progress but our view is that these topics are now a core part of the database curriculum. The final comment comes from Michael Rappa of the Institute for Advanced Analytics: "Big data isn't a new speciality or suite of tools we have to train people into, as much as it's a new organizational reality that everyone will need to adjust to occupationally" (Rappa, 2013)

References

Anderson P., Bowring J., McCauley R., Pothering G. & Starrr C. (2014) *An Undergraduate Degree in Data Science: Curriculum and a Decade of Implementation Experience SIGCSE'14*, March 5–8, 2014, Atlanta, Georgia, USA.

Antony J., (2013) "What does the future hold for quality professionals in organisations of the twenty-first century?", The TQM Journal, Vol. 25 Iss: 6, pp.677 – 685

Barnes N.D. (2014) *Analyse this: the Big Demand for Big Data Professionals* Information Management Jan/Feb 2014 pp 34 -37

Boyd D., Crawford K. (2012) *Critical Questions for Big Data* Information, Communication and Society 15:5, 662-679

Demchenko, Y., Grosso, P., De Laat, C. & Membrey, P. (2013). Addressing big data issues in Scientific Data Infrastructure. In: 2013 International Conference on Collaboration

Gartner http://www.datanami.com/2014/03/13/hadoop_and_nosql_now_data_warehouse-worthy_gartner/ accessed 11 June 2014

Grossniklaus M., Maier D., (2012) *The Curriculum forecast for Portland: Cloudy with a chance of Data* ACM Sigmod Record 41: 1 pp 74 -77

Java Developers Journal 2004 http://www2.sys-con.com/itsg/virtualcd/java/archives/0410/scott/index.html accessed 11 June 2014

Jelinek M, Bergey P. (2013) *Innovation as the strategic driver of sustainability: big data knowledge for profit and survival* IEE Engineering Management Review 41: 2

Kambatla K., Kollias G., Kumar V., Grama A. (2014) *Trends in Big Data Analytics* Journal of Parallel Distributed Computing 74 pp 2561-2573

Kraska T. (2013) *Finding the Needle in the Big Data Systems Haystack* IEEE Internet Computing pp 84 -86

Leavitt Communications n.d. Whatever Happened to Object Oriented Databases http://www.leavcom.com/db_08_00.htm accessed 17 June 2014

Madden S. (2012) *From Databases to Big Data* IEEE Internet Computing pp 4-6

Market Research Media http://www.marketresearchmedia.com/?p=568 accessed 11 June 2014

Patel P. (2013) *A Degree in Big Data* IEEE Spectrum *spectrum.ieee.org/at-work/tech-careers/a-degree-in-big-data* accessed 12 June 2014

Quality Assurance Agency for Higher Education (QAA) Computing Benchmark 2007 http://www.qaa.ac.uk/Publications/InformationAndGuidance/Documents/computing07.pdf

Rappa M. (2012) cited in Thibodeau P. (2013) *Big Data Bring Big Academic Opportunities* ComputerWorld 21 09 2012 http://www.computerworld.com/s/article/9231523/Big_data_brings_big_academic_opportunities accessed 15 June 2014

Sattar A., Lorenzen T. & Nallamaddi K. (2013) *Incorporating NoSQL into a Database Course* ACM Inroads 4:2 pp. 50 -53

Sherestha R. (2014) *Big Data and Cloud Computing* Applied Radiology 43: 3 pp 32-34

Stanier C. (2012) *Introducing NoSQL into the Database Curriculum* TLAD 2012 pp. 61 -72

Stonebraker M.(2013) *Big Data is 'Buzzword' du Jour'* Comm. of the ACM 56:9 pp 10-11