**Data Science Lab Innovations**

**Assignment Cover Sheet**

**Module code:** COMM420DAJ 1OJ21

**Module name:** Data Science Professional   
 Practice 1

**Student number:**  720027410

**Candidate number:**  212204

**Word count:** 2971

**Coursework 2:** Detailed Plan for a new Corporate Continuing Professional Development   
 event

Table of Contents

1. Abstract……………………………………………………………05
2. Executive Summary……………………………………………….05
3. Introduction…………………………………………………….….06
4. Phases in the Processing Pipeline……………………………….....07
5. Data Acquisition and Recording………………...………………....07
6. Information Extraction and Cleaning………………...…………….07
7. Data Integration, Aggregation, and Representation………………..08
8. Query Processing, Data Modelling, and Analysis………………....08
9. Interpretation………………...………………...……………….......08
10. Heterogeneity and Incompleteness………………...………………09
11. Timeliness………………...………………...………………...……09
12. Privacy………………...………………...………………...……….09
13. Human Collaboration………………...………………...…………..10
14. System Architecture………………...………………...……………10
15. Conclusion………………...……….………………...………….….11
16. Bibliography………………...……....………………...……………12
17. References………………………...………………...……………...13

**What you’ll learn and how you can apply it:**

* SOLID principles of Big Data Analytics
* Clean Code basics
* Ensure that dependencies are well managed so that the code remains flexible, robust, and reusable
* Understand costs and trade-offs associated with object-oriented design and get a leg up in ensuring that your code base is much cleaner
* How architecture risk can impact the success of a system
* Techniques and patterns for addressing risk areas related to security, elasticity, and data integrity
* Navigate and use the Hadoop Distributed File Systems (HDFS)
* Professional practices

**Recommended Preparation:**

[Clean Architecture](https://www.oreilly.com/library/view/clean-architecture-a/9780134494272/part3.xhtml) (book) - Part III Design Principles  
[Clean Code Video Series](https://www.oreilly.com/library/view/clean-code/9780134661742/) (video)  
[Clean Coder Video Series](https://www.oreilly.com/library/view/clean-coder-clean/9780134843803/) (Expanded Edition)  
[Clean Code](https://www.oreilly.com/library/view/clean-code/9780136083238/) (book)

**Prerequisites:**

* Software development experience
* Familiarity of common software design

**Schedule:**

* Segment 1:   
  Introduction (15 min)
* Segment 2:   
  Phases in the Processing Pipeline (55 min)

1. Data Acquisition and Recording (8 min)
2. Information Extraction and Cleaning (7 min)
3. Data Integration, Aggregation, and Representation (5 min)
4. Query Processing, Data Modelling, and Analysis (3 min)
5. Interpretation (2 min)

*Break (length: 10 min)*

1. Heterogeneity and Incompleteness (5 min)
2. Timeliness (6 min)
3. Privacy (4 min)
4. Human Collaboration (4 min)
5. System Architecture (6 min)

* Segment 3:   
  Conclusion and Q&A (20 min)

1. **Abstract**

In this world of information, the term BIG DATA has emerged with new opportunities and challenges to deal with the massive amount of data. BIG DATA has earned a place of great importance and is becoming the choice for new research. To find the useful information from massive amount of data to organizations, we need to analyse the data. Mastery of data analysis is required to get the information from unstructured data on the web in the form of texts, images, videos, or social media posts. This document presents an overview on Big Data, Advantages, and its scope for the future research. Big Data present opportunities as well as challenges to the researchers. An overview on opportunities to fintech, healthcare, technology etc. is given.

***Keywords*** - big *data; Hadoop; Map Reduce; HDFS; data mining.*

1. **Executive Summary**

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of Big Data. While the promise of Big Data is real for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 there is currently a wide gap between its potential and its realization.

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modelling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analysed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

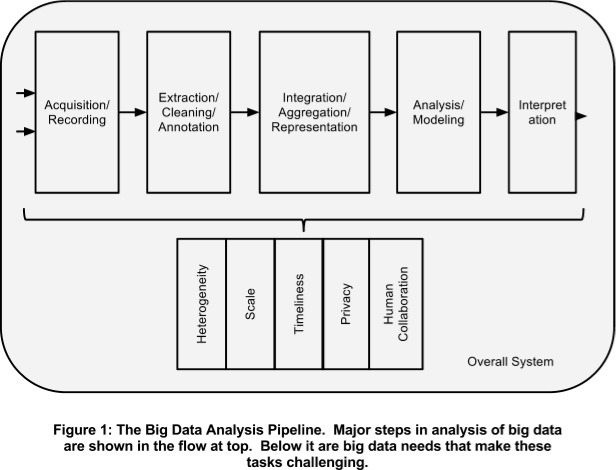
1. **Introduction**

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

Scientific research has been revolutionized by Big Data [CCC2011a]. The Sloan Digital Sky Survey [SDSS2008] has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer’s job to one where the pictures are all in a database already and the astronomer’s task is to find interesting objects and phenomena in the database. In the biological sciences, there is now a well- established tradition of depositing scientific data into a public repository, and of creating public databases for use by other scientists. In fact, there is an entire discipline of bioinformatics that is largely devoted to the curation and analysis of such data. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially.

Big Data has the potential to revolutionize not just research, but also education [CCC2011b]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [DF2011]. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. There is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance. [1]

It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality [CCC2011c], by making care more preventive and personalized and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates [McK2011] a savings of 300 billion dollars every year in the US alone.



1. **Phases in the Processing Pipeline**
2. **Data Acquisition and Recording:**

Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. For example, suppose one sensor reading differs substantially from the rest: it is likely to be due to the sensor being faulty, but how can we be sure that it is not an artifact that deserves attention? In addition, the data collected by these sensors most often are spatially and temporally correlated (e.g., traffic sensors on the same road segment). We need research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not missing the needle in the haystack. Furthermore, we require “on-line” analysis techniques that can process such streaming data on the fly, since we cannot afford to store first and reduce afterward.

1. **Information Extraction and Cleaning:**Frequently, the information collected will not be in a format ready for analysis. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements (possibly with some associated uncertainty), and image data such as x-rays. We cannot leave the data in this form and still effectively analyse it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge.
2. **Data Integration, Aggregation, and Representation:**Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this must happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then “robotically” resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution. In August, 2013 the definition was further enhanced to include, "veracity, variability, visualization, and value" which gave a newer perspective to it.[2]
3. **Query Processing, Data Modelling, and Analysis:**Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. As noted previously, real-life medical records have errors, are heterogeneous, and frequently are distributed across multiple systems.
4. **Interpretation:**Having the ability to analyse Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, must interpret these results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we saw above, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the computer. The computer system must make it easy for her to do so. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in. The recent mortgage-related shock to the financial system dramatically underscored the need for such decision-maker diligence -- rather than accept the stated solvency of a financial institution at face value, a decision-maker must examine critically the many assumptions at multiple stages of analysis.
5. **Heterogeneity and Incompleteness:**When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stays, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems.

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge. Recent work on managing probabilistic data suggests one way to make progress. Government Sector being the highest among them, offers wide range of opportunities for bigdata analyst and researchers. [3]

1. **Timeliness:**The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyse. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data.   
     
   There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed - potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user’s purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination. [4]

1. **Privacy:**The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data. [5]
2. **Human Collaboration:**  
   Despite the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Indeed, CAPTCHAs precisely exploit this fact to tell human web users apart from computer programs. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modelling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

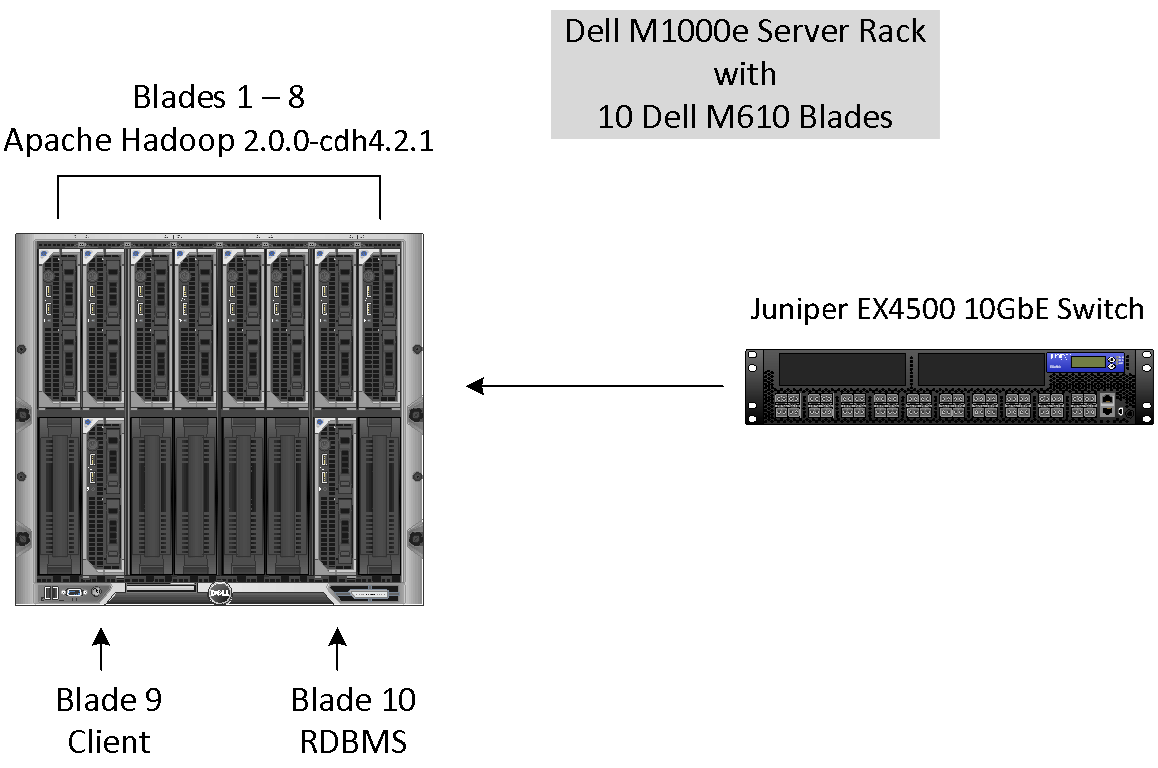
In today’s complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system must accept this distributed expert input and support their collaboration. [6]

1. **System Architecture:**  
   Companies today already use, and appreciate the value of, business intelligence. Business data is analysed for many purposes: a company may perform system log analytics and social media analytics for risk assessment, customer retention, brand management, and so on. Typically, such varied tasks have been handled by separate systems, even if each system includes common steps of information extraction, data cleaning, relational-like processing (joins, group-by, aggregation), statistical and predictive modelling, and appropriate exploration and visualization tools. [7]

With Big Data, the use of separate systems in this fashion becomes prohibitively expensive given the large size of the data sets. The expense is due not only to the cost of the systems themselves, but also the time to load the data into multiple systems. In consequence, Big Data has made it necessary to run heterogeneous workloads on a single infrastructure that is sufficiently flexible to handle all these workloads. The challenge here is not to build a system that is ideally suited for all processing tasks. Instead, the need is for the underlying system architecture to be flexible enough that the components built on top of it for expressing the various kinds of processing tasks can tune it to efficiently run these different workloads.

If users are to compose and build complex analytical pipelines over Big Data, it is essential that they have appropriate high-level primitives to specify their needs in such flexible systems. The Map- Reduce framework has been tremendously valuable but is only a first step. Even declarative languages that exploit it, such as Pig Latin, are at a rather low level when it comes to complex analysis tasks. Similar declarative specifications are required at higher levels to meet the programmability and composition needs of these analysis pipelines. Besides the basic technical need, there is a strong business imperative as well.

**Blade configuration:**   
Intel Xeon X5667 3.07GHz processor   
Dell PERC H700 Integrated RAID controller   
Disk size: 543 GB   
FreeBSD iSCSI Initiator driver   
HP P2000 G3 iSCSI dual controller   
Memory: 94.4 GB   
Linux 2.6.32 [LP+2009] [8]



1. **Conclusion**We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this document which we must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone.

Due to the gargantuan increase in the amount of data in various fields, it becomes a major challenge to handle the data efficiently. Thus to come up with possible solutions to these challenges one needs to understand the concept of big data, its handling methodologies and furthermore improve the approaches in analysing big data. With the advent of social media, the need for handling big data has increased monumentally. If Facebook, WhatsApp, Twitter produce data which keeps increasing exponentially every year (or a few years) then handling such huge data is something to be efficiently dealt with.

1. **Bibliography**

[CCC2011a] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.

[CCC2011b] Advancing Personalized Education. Computing Community Consortium. Spring 2011.

[CCC2011c] Smart Health and Wellbeing. Computing Community Consortium. Spring 2011.

[SDSS2008] SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, available at http://www.sdss3.org/collaboration/description.pdf

[DF2011] Getting Beneath the Veil of Effective Schools: Evidence from New York City. Will Dobbie, Roland G. Fryer, Jr. NBER Working Paper No. 17632. Issued Dec. 2011.

[McK2011] Big data: The next frontier for innovation , competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[LP+2009] Computational Social Science. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer,Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Science 6 February 2009: 323 (5915), 721-723.

[K2.6 & B1.6] Provides an example of when they have led a process leading to the achievement of   
an organisational objective and how their project management skills had a positive impact on quality and cost

[K7.1] Provides an example of where they have used market analysis tools (SWOT) including decisions made in terms of value for money

[K8.3] (Distinction): Describes an example of when they have coached or mentored colleagues, peers or team members and identifies the benefits of this

[S5.4] (Distinction): Can describe examples of when they have adapted scientific strategy or delivery to consistently meet requirements. e.g. client, regulatory, ethical, geographic

1. **References**

[1] Douglas, Laney. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

[2] Blog post: Mark van Rijmenam titled "Why the 3V's Are Not Sufficient to Describe Big Data".

[3] Jean Yan, April 9, 2013 ―Big Data, Bigger Opportunities Bowman, M. Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.

[4] ―Research in Big Data and Analytics: An Overview‖ International Journal of Computer Applications (0975 – 8887) Volume 108 – No 14, December 2014

[5] Blog post: Thoran Rodrigues in Big Data Analytics, titled ―10 emerging technologies for Big Data‖, December 4, 2012

[6] 2013 IEEE 37th Annual Computer Software and Applications Conference. Elisa Bertino Cyber Center, CERIAS and CS Department Purdue University West Lafayette, Indiana (USA) ―Big Data - Opportunities and Challenges Panel Position Paper

[7] Wie, Jiang, Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core

[5] Big Data- The definitive guide to the revolution in business analytics-Fujitsu

[6] Khttp://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System -HDFS

[7] http://searchcloudcomputing.techtarget.com/definiti on/MapReduce

[8] http://www.informit.com/articles/article.aspx?p=2008905