Module 6: Big Data Management

Stage				1			
Semester				2			
Module Title				Big Data Management			
Module Number				6			
Module Status				Mandatory			
Module ECTS Credits				10			
Module NFQ level				9			
Pre-Requisite Module Titles				None			
Co-Requisite Module Titles				None			
Capstone Module				No			
List of Module Teaching Personnel				Ms Jennifer Treanor			
Contact Hours				**Non-contact Hours**			**Total Effort (hours)**
72				**128**			**200**
Lecture	**Practical**	**Tutorial**	**Seminar**	**Assignment**	**Placement**	**Independent Work**	
36	36			60		68	
Allocation of Marks (Within the Module)							
	Continuous Assessment	**Project**	**Practical**	**Final Examination**			**Total**
Percentage Contribution	**60**			**40**			**100**

Intended Module Learning Outcomes

On successful completion of this module the learner will be able to:

1. Critically analyse the differences between traditional data stores and Big Data datasets
2. Critique the consequences of the main failure points of traditional systems with regard to this level of data
3. Describe in detail the lambda architecture and how it is implemented via current technologies
4. Convert RDB-style data to fact-based data structures
5. Implement and use a distributed file system for storage of Big Data – batch layer
6. Connect a NoSQL database to a master dataset for basic querying - serving layer
7. Implement a real-time database to handle smaller data quantities with higher input frequencies
8. Connect the three layers together on a real-world data set

Module Objectives

This module aims to equip the learner with the skills to implement, from the batch to the speed layer, an end-to-end Big Data storage system using the most current technologies. As a grounding to the subject area, the learner will be guided through an overview of the traditional approach of data storage and access, with all theory grounded in real-world technological examples. As technologies have progressed, the availability of data has increased dramatically. The volumes of data dealt with in modern systems are far beyond what traditional systems can handle. During this module, the main failure points of traditional systems with regard to this level of data will be explored. Each layer of the Lambda Architecture will be explored in detail from theory through to implementation via current technologies. At the lowest layer, the module will demonstrate how to store Big Data in the fact-based model in a distributed file system, namely Hadoop Distributed File System (HDFS). This layer is then connected to a read-oriented database, such as MongoDB or ElephantDB, depending on the data type stored, to create the Serving Layer of the Lambda Architecture. Finally, this will be connected to a light-weight database that can handle high-volume reads and writes to implement the high-level Speed Layer of the Lambda Architecture. All practical work will be done on real-world data to emphasise the need for Big Data systems.

Module Curriculum

- **Overview of Traditional Approach**
  Distributed Databases: distributed data storage, transaction control, commit protocols, concurrency control
  Data Warehousing: architectures, data marts, top-down/bottom-up methodologies
  Scaling Problems

- **Introduction to Big Data**
  Lambda Architecture overview, physical requirements, data storage: raw data, fact-based model, data sources, NoSQL

- **Batch Layer**
  Master dataset: Storage requirements, operations, chunking and replication, name nodes and data nodes
  Recomputation Algorithms: recomputation vs incremental algorithms, hadoop MapReduce
  Distributed File Systems: Hadoop – HDFS, operations, implementation

- **Serving Layer**
  Storage Requirements, importing batch data, performing queries on batch data, Serving Layer databases MongoDB, ElephantDB

- **Speed Layer**
  Real-time results, mapping speed-layer results to serving layer results, merging algorithms, latency

- **Case Studies**
  Google F1 DRDBMS, Implement distributed RDB across multiple data stores, Twitter Streaming API, JSON data strucutres

Reading Lists and other learning materials

Recommended Reading

Marz N et al., 2014, *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*

Gates A, 2011 *Programming Pig*

Secondary Reading

Holmes A, 2012, *Hadoop in Practice*

Perera S, 2013, *Hadoop MapReduce Cookbook,*

Kyle Banker, 2011, *MongoDB in Action,*

Module Learning Environment

Accommodation

Lectures are carried out in class rooms / lecture halls in the College. Lab tutorials are carried out in computer labs throughout the Campus. All have the software required to deliver the programme.

Library

All learners have access to an extensive range of physical and electronic (remotely accessible) library resources. The library monitors and updates its resources on an on-going basis, in line with the College's Library Acquisition Policy. Lecturers update reading lists for this course on an annual basis as is the norm with all courses run by Griffith College.

Module Teaching and Learning Strategy

Each week involves both classes and practical laboratory sessions

Classes are used to deliver theoretical content and may be supported by online delivery of notes, examples, and web resources.

Laboratory Practicals are used to provide continuous progression of theory presented in lectures with each session building upon ideas of the previous lectures and laboratory sessions. Use of multiple data nodes, i.e. servers is vital to the understanding of the complexities of Big Data. Distributed data sets will be used from

the early stages of the course.

Module Assessment Strategy

Continuous assessment is based on a combination of some of the following:

- Programming Assignments
- Report Writing
- Oral Examination

Element	Weighting	Type	Description	Learning Outcomes Assessed
1	10%	Programming Assignment	Implement batch layer with existing technologies for a given dataset, along with MapReduce calculations	1, 4, 5
2	20%	Programming Assignment	Implement Serving Layer with existing technologies and connect to a Batch Layer	5,6
3	30%	Programming Assignment	Connect Batch Layer, Serving Layer and Speed Layer using existing technologies	6,7,8
3	40%	Closed book exam	Implement informative visual analysis of a large dataset	1,2,3