Subject Outline

Subject Name	Big Data Management and Processing
Subject Code	MA5831
Study Period	SP85 2020
Study Mode	External
Campus	JCU online
Subject Coordinator	Neil Fraser

We acknowledge the Traditional Owners of the lands and waters where our University is located and actively seek to contribute and support the JCU Reconciliation Statement, which exemplifies respect for Australian Aboriginal and Torres Strait cultures, heritage, knowledge and the valuing of justice and equity for all Australians.

Cairns
Singapore
Townsville

Pre-requisites

Before attending this course, you should have experience in writing SAS programs or working with other languages such as R.

Although it is not compulsory, it is recommended for students to complete **CP5804 Database Systems** before enrolling in this subject.

Students should also have an understanding of the basics of computing, data management, different data types and some programming experience in SAS and SQL.

Subject outline preparation

This subject outline has been prepared by [Type here] for the College of [Type here], Division of [Type here], James Cook University. Updated [Type here].

Q1. This subject is offered across more than one campus and/or mode and/or teaching period within the one calendar year.	Yes ☒	No ☐
Q2. If yes (Q1), the design of all offerings of this subject ensure the same learning outcomes and assessment types and weightings.	Yes ☒	No ☐
Q3. If no (Q2), [Type here] has authorised any variations, in terms of equivalence.		

Subject outline peer reviewer

Name	Ron White
Position	Professor
Date reviewed	

Staff contact details

Teaching team	**Staff member**	**Room**	**Phone**	**Email**	**Consultation times***
Subject Coordinator	Neil Fraser	JCU Cairns		Neil.fraser@jcu.edu.au	Weekdays Business Hours
Tutor 1	Banmali Pradhan			banmali.pradhan@jcu.edu.au	Online in unit discussion boards
Tutor 2	Mostafa Shaikh			mostafa.shaikh@jcu.edu.au	Online in unit

Teaching team	Staff member	Room	Phone	Email	Consultation times*
					discussion boards

*Other consultation times by appointment only.

Note:

For email enquiries, **please type MA5831 in the subject line**, and ensure your email is sent from your JCU email account. Should the matter be urgent or not responded to by the subject co-ordinator in a timely fashion, please contact the school office on 4781 4785. **Do not email assessment to the lecturer unless prior approval has been granted.**

Contents

1 Subject at a glance

1.1 Student participation requirements

The JCU Learning, Teaching and Assessment Policy (4.3) indicates that, "a **3 credit point subject** will require a **130 hour work load** of study-related participation including class attendance over the duration of the study period, **irrespective of mode of delivery**". This work load comprises **timetabled hours** and **other attendance requirements**, as well as **personal study hours,** including completion of online learning activities and assessment requirements. Note that "attendance at specified classes will be a mandatory requirement for satisfactory completion of some subjects" (Learning, Teaching and Assessment Policy, 5.10); and that additional hours may be required per week for those students in need of **English language, numeracy** or **other learning support.**

Key subject activities	**Time**	**Day and date**	**Room/Location**
Online Collaboration session	20:00-21:00	Thursday Week 1-6	LearnJCU

For information regarding class registration, visit the Class Registration Schedule.

1.2 Key dates

Key dates	**Date**
Census date	See 2020 Study Period and Census Dates
Last date to withdraw without academic penalty	See 2020 Study Period and Census Dates
Assessment task 1: Quiz[10 %]	Due Wednesday 11:59 pm of Week 2
Assessment task 2: SAS data quality profiling and standardizing Case [20 %]	Due Sunday 11:59 pm of Week 4
Assessment task 3: Literature review [40 %]	Due Wednesday 11:59 pm of Week 6
Assessment task 4: SAS Managing data with HIVE and PIG in HADOOP case study [20 %]	Due Wednesday 11:59pm of Week 7
Assessment task 5: SAS Course chapter completion [10 %]	Due Wednesday 11:59pm of Week 7

2 Subject details

2.1 Subject description

This subject will provide students with cutting-edge tools and techniques for high-performance and large-scale computing, with focus on computer models and software designed to handle Big Data sets in a distributed and/or parallel fashion. Particular focus will be given to distributed and parallel computing using Map-Reduce/Hadoop and similar models for processing Big Data sets.

2.2 Subject learning outcomes

Students who successfully complete this subject will be able to:

1. Compare and evaluate different systems and approaches for high-performance and large-scale computing for analytics for standard data and big data
2. Manage and prepare data using standard management frameworks for the purpose of transforming, cleaning to ensuring classical characteristic outcomes are achieved
3. Perform data management tasks to improve data quality, entity resolution and data monitoring
4. Examine and deploy data processing tasks in the Hadoop ecosystem for big data and critically evaluate the combination of Hadoop and SAS to overcome big data challenges
5. Choose and apply different techniques and software for distributed and cloud computing of big data
6. Conduct a written review of a current data processing technology to establish a critical baseline understanding of the academic research with regards to a new trend

This subjects is SAS accredited subject therefore we will be using SAS exclusively.

Explicit SAS modules:

i. Big Data Challenges and Analysis Driven Data (4 hours)
ii. Preparing Data for Analysis and Reporting (28 hours)
iii. Introduction to Hadoop and SAS
iv. Hadoop Data Management with Hive, Pig and SAS

These outcomes will contribute to your overall achievement of **course learning outcomes.** Your course learning outcomes can be located in the entry for your course in the electronic JCU Course and Subject Handbook 2020 (click on 'Course Information' bar/ select 'Undergraduate Courses' or 'Postgraduate Courses'/ select relevant course/ scroll down to 'Academic Requirements for Course Completion', 'Course learning outcomes').

2.3 Learning and teaching in this subject

In the first three weeks the subject, we will cover **preparing standard data** for data science teams involved in analysis, reporting and model development. This will follow the SAS Data Science academy curriculum in "Preparing Data for Analysis and reporting"

In the last three weeks of the subject, we will cover **preparing big data** for data science teams involved in analysis, reporting and model development. This will follow the SAS Data Science academy curriculum in "Hadoop Data Management with Hive, Pig and SAS "

2.4 Student feedback on subject and teaching

As part of our commitment at JCU to improving the quality of our courses and teaching, we regularly seek feedback on your learning experiences. Student feedback informs evaluation of subject and teaching strengths and areas that may need refinement or change. ***YourJCU Subject and Teaching Surveys*** provide a formal and confidential method for you to provide feedback about your subjects and the staff members teaching within them. These surveys are available to all students through LearnJCU. You will receive an email invitation when the survey opens. We value your feedback and ask that you provide constructive feedback about your learning experiences for each of your subjects, in accordance with responsibilities outlined in the Student Code of Conduct. Refrain from providing personal feedback on topics that do not affect your learning experiences. Malicious comments about staff are deemed unacceptable by the University.

2.5 Subject resources and special requirements

This subject will use e-learning, video tutorials, documentation and webinars based around the SAS Academy of Data Sciences courses.

The following main textbook is required for the course and is available online through the JCU library.

Data Warehousing Fundamentals for IT Professionals, Second Edition, Wiley, Paulraj Ponniah

Additionally, you can find the most appropriate library subject resources, including dedicated discipline libguides, relevant databases and access to library services and staff through the *Your Library* tool, in your LearnJCU subject site.

3 Assessment details

3.1 Requirements for successful completion of subject

In order to pass this subject, you must:

- Attain an aggregate score of 50% or higher across all assessed elements.
- Submit only original, non-plagiarised work for assessment.

Assessment items and final grades will be reviewed through moderation processes (Learning, Teaching and Assessment Policy, 5.13-5.18). It is important to be aware that assessment “is always subject to final ratification following the examination period and that no single result represents a final grade in a subject” (Learning, Teaching and Assessment Policy, 5.22.).

3.1.1 Inherent requirements

Inherent requirements are the fundamental abilities, attributes, skills and behaviours needed to achieve the learning outcomes of a course while preserving the academic integrity of the university’s learning, assessment and accreditation processes. Students and prospective students must be able to demonstrate that they have acquired or have the ability to acquire the inherent requirements for their degree.

Reasonable adjustments may be made to assist students manage additional circumstances impacting on their studies provided these do not change the academic integrity of a degree. Reasonable adjustments do not alter the need to be able to demonstrate the inherent requirements of the course. Students who believe they will experience challenges completing their degree or course because of their disability, health condition or other reason should discuss their concerns with an AccessAbility Services team member or a member of College staff, such as the Course Coordinator. In the case where it is determined that inherent requirements cannot be met with reasonable adjustments, the University staff can provide guidance regarding other study options.

3.2 Feedback on student learning

Feedback for students will be provided on some assessment items.

3.3 Assessment tasks

ASSESSMENT TASK 1: ONLINE QUIZ

Aligned subject learning outcomes	• Compare and evaluate different systems and approaches for high-performance and large-scale computing for analytics for standard data and big data • Manage and prepare data using standard management frameworks for the purpose of transforming, cleaning to ensuring classical characteristic outcomes are achieved
Aligned professional standards/ competencies	
Group or individual	Individual
Weighting	10%
Due date	Wednesday 11:59 Week 2

ASSESSMENT TASK 1: DESCRIPTION

Assignment 1 will be in the following answer formats

- Multiple choice quiz

ASSESSMENT TASK 1: CRITERIA SHEET

All assessment questions will be marked by comparing student answers to a model set of solutions and marking scheme prepared by the lecturer.

ASSESSMENT TASK 2: DATA QUALITY PROFILING AND STANDARDIZING CASE STUDY

Aligned subject learning outcomes	Perform data management tasks to improve data quality, entity resolution and data monitoring
Aligned professional standards/ competencies	SAS Big Data Preparation, Statistics, and Visual Exploration
Group or individual	Individual
Weighting	20%
Due date	11:59 PM AEST Sunday of Week 4

ASSESSMENT TASK 2: DESCRIPTION

Assignment 2 will be in the following answer formats

- Written assessment: Case study on Preparing Data for Analysis and Reporting

ASSESSMENT TASK 2: CRITERIA SHEET

All assessment questions will be marked using a marking rubric on the subject’s website

ASSESSMENT TASK 3 : DATA AND LITERATURE REVIEW

Aligned subject learning outcomes	Conduct a written review of a current data processing technology to establish a critical baseline understanding of the academic research with regards to a new trend
Aligned professional standards/ competencies	Academic Literature Review
Group or individual	Individual
Weighting	40%
Due date	11:59 PM AEST Wednesday of Week 6

ASSESSMENT TASK 3: DESCRIPTION

Assignment 3

- Written assessment: Literature review

ASSESSMENT TASK 3: CRITERIA SHEET

All assessment questions will be marked using a marking rubric on the subject's website

ASSESSMENT TASK 4: MANAGING DATA WITH HIVE AND PIG IN HADOOP CASE STUDY

Aligned subject learning outcomes	Examine and deploy data processing tasks in the Hadoop ecosystem for big data and critically evaluate the combination of Hadoop and SAS to overcome big data challenges Choose and apply different techniques and software for distributed and cloud computing of big data
Aligned professional standards/ competencies	Certified Big Data Professional Using SAS® Part 1 SAS Big Data Programming and Loading
Weighting	20%
Date	11:59 PM AEST Wednesday of Week 7
See Special Consideration, Supplementary, Deferred and Special Examinations Policy	

ASSESSMENT TASK 4: DESCRIPTION

Assignment 4 will be in the following answer formats

- Written Assessment : Case Study and Review: Hadoop Data Management with Hive, Pig, and SAS®

ASSESSMENT TASK 4: CRITERIA SHEET

All assessment questions will be marked using a marking rubric on the subject's website

ASSESSMENT TASK 5: SAS CHAPTER COMPLETION

Aligned subject learning outcomes	SAS Academy of Data Science course Chapter completion
Aligned professional standards/ competencies	Certified Big Data Professional Using SAS® Part 1 SAS Big Data Programming and Loading
Weighting	10%
Date	11:59 PM AEST Wednesday of Week 7
See Special Consideration, Supplementary, Deferred and Special Examinations Policy	

ASSESSMENT TASK 5: DESCRIPTION

Assignment 5 requires the completion of all related SAS Academy of Data Science course chapters covered in this subject. These chapters and related activities have already been completed as part of the weekly content and progress has been automatically monitored in the SAS Academy of Data Science for assessment and marking, and no further submission or action is required in Learn JCU.

ASSESSMENT TASK 3: CRITERIA SHEET

All course chapter completion will be automatically assessed using SAS Academy.

4 Submission and return of assessment

4.1 Submission of assessment

All assessments are submitted through Learn Ultra.

Note that the Learning, Teaching and Assessment Policy (5.22.3) outlines a uniform formula of penalties that will be imposed for submission of an assessment task after the due date. **This formula is 5% of the total possible marks for the assessment item per day including part-days, weekends and public holidays**. After 20 days, the assessment item thus would be awarded 0 marks (i.e. 5% x 20 = 100% of total possible marks in penalties) and would not be marked.

4.2 Return of assessment

Feedback on marked assessments will be available in the Gradebook in Learn Ultra.

Due to short turnaround time, the feedback can potentially be brief. Please communicate with subject coordinator if the detailed feedback is required.

It is the responsibility of students to view their marks for each within-session assessment on Learn JCU within 20 working days of posting. If there are any discrepancies, students must contact the unit convenor immediately. Failure to do so will mean that queries received after the release of final results regarding assessment marks (not including the final exam mark) will not be addressed.

Please see The Learning Centre website for other important student information pertaining to plagiarism and referencing, examinations advice and student support services.

Please see the Current Students web page for links to all student resources and support services to optimise your academic and personal success.

Please see the Learn Student Guide web page for general advice on plagiarism, referencing and examinations. Here, you can also access individual and group assessment task cover sheets. Note that cover sheets are only required for hard copy submissions.

5 Subject calendar

Please note, the sequence of some topics may change due to staff availability, resourcing, or due to unforeseen circumstances.

Week/Date/Module		Topics	Tutorial/Activities	Readings/Preparation	Learning outcome/Assessment
O	Orientation	SAS Academy of Data Sciences Start Course: Preparing Data for Analysis and Reporting Tutorial: Dataflux Data management Studio Chapter 1 Course Flow	Start Free E-learning SAS Programming 1: Essentials	This subject will use E-learning, Video Tutorials, Documentation and Webinars based around the SAS Academy of Data Sciences	
1	Foundations and Principles in Data processing for Analysis and Data Science	Topic 1: Foundations of Standard digital data and Big Data Topic 2: Data Processing in Enterprises Topic 3: Foundations of data processing techniques Topic 4: Foundations of data storage technologies	Self-Learning Practical: Big data challenges and analysis-driven data Continue Free E-learning SAS Programming 1: Essentials	Data Warehousing Fundamentals for IT Professionals, Second Edition, Wiley, Paulraj Ponniah (Chapter 2). Data lakes in business intelligence: reporting from the trenches by Llave, Marilex Rea Procedia Computer Science, 2018, Volume 138	Learning Objective 1 A1 Quiz
2	Data processing: Characteristics, constraints and quality	Topic 1: Common Causes of poor Data quality Topic 2: Key characteristics of transformed data	SAS Academy of Data Sciences > Big Data Preparation, Statistics, and Visual Exploration	Chapter 12 Data Warehousing Fundamentals for IT Professionals, Second Edition, Wiley, Paulraj	Learning Objective 2 A2 Data Quality profiling and standardizing Case Study

Week/Date/Module		Topics	Tutorial/Activities	Readings/Preparation	Learning outcome/Assessment
	considerations	Topic 3: Working with standard data flows Topic 4: Data processing constraints	Preparing Data for Analysis and Reporting Self Learning Practical 1: Planning (PDAR :Chapter 3) Self Learning Practical 2: Acting (PDAR :Chapter 4) Self Learning Practical 3: Monitoring (PDAR :Chapter 5) Continue Free E-learning SAS Programming 1: Essentials	Ponniah (available online). SAS Course Notes: Preparing Data for Analysis and Reporting	
3	Parsing, matching and standardising data	Topic 1 The standard data-cleansing process steps Topic 2 Entity resolution with multiple inputs and outputs Topic 3 SAS Quality Knowledge base (QVB) and customization Topic 4 Data Types and Definitions Topic 5: Literature Reviews	<SAS Academy of Data Sciences> Big Data Preparation, Statistics, and Visual Exploration Preparing Data for Analysis and Reporting Self-Learning Practical 1: Self-learning Practical 1: Parse Match and Standardize Free E-learning SAS Programming 1: Essentials	Chapter 13 Data Warehousing Fundamentals for IT Professionals, Second Edition, Wiley, Paulraj Ponniah (available online). SAS Course Notes: Preparing Data for Analysis and Reporting	Learning Objective 3 and 6 A3 Literature Review (due Week 6)

Week/Date/Module		Topics	Tutorial/Activities	Readings/Preparation	Learning outcome/Assessment
4	Big Data Processing	Topic 1: Big data processing and storage Topic 2: Hadoop ecosystem essentials Topic 3: HDFS operations Topic 4: How Hadoop processes Data	<SAS Academy of Data Sciences > Big Data Preparation, Statistics, and Visual Exploration Introduction to SAS® and Hadoop Self-Learning Practical 1 Introduction to SAS® and Hadoop Self-Learning Practical 2: SAS® and Hadoop data operations Free E-learning SAS Programming 1: Essentials	A big data methodology for categorising technical support requests using Hadoop and Mahout by Duque Barrachina, Arantxa; O'Driscoll, Aisling Journal of Big Data, 12/2014, Volume 1, Issue 1 Data-intensive text processing with MapReduce by Lin, Jimmy; Dyer, Chris 2010, Chapter 2	Learning Objective 4
5	Hadoop essentials: data connectors, resource management and advanced Hive usage	Topic 1: Data connectors for Hadoop Topic 2: Hadoop and YARN Topic 3: Hive and HiveQL	< SAS Academy of Data Sciences > Big Data Preparation, Statistics, and Visual Exploration >Hadoop Data Management With Hive, Pig and SAS. Warm-Up Activity: The Apache Hadoop Project Self-Learning Practical 1: Hadoop Essentials: Hue interface	Data-intensive text processing with MapReduce by Lin, Jimmy; Dyer, Chris 2010, Chapter 3	Learning Objective 5 A4: Managing Data with Hive and Pig in Hadoop case Study

Week/Date/Module		Topics	Tutorial/Activities	Readings/Preparation	Learning outcome/Assessment
			and Sqoop Self-Learning Practical 2: HIVE and Hive QL Free E-learning SAS Programming 1: Essentials		
6	Hadoop data management with Hive, Pig and SAS	Topic 1: Pig and Pig Latin Topic 2: In-Memory Data Processing with SAS In-Memory statistics	<SAS Academy of Data Sciences > Big Data Preparation, Statistics, and Visual Exploration Self-Learning Practical 1: Hadoop data management with Hive, Pig and SAS Self-Learning Practical 2: Getting started Self-Learning Practical 3: SAS, In-Memory and Hadoop		Learning Objective 5 Assessment 5: SAS CHAPTER COMPLETION