DEGREE: MSc Data Analytics

Module: Big Data Analytics

Assignment Title: Cloud-Based Big Data Analytics with Apache Spark and Hadoop Ecosystem

Assignment Type: Report

Word Limit: 3000 words (+/- 300)

Weighting: 100%

Issue Date: 19/11/2024

Submission Date: 05/02/2025

Feedback Date: 26/02/2025

Plagiarism:

When submitting work for assessment, students should be aware of the InterActive/Canvas guidance and regulations concerning plagiarism. All submissions should be your own, original work.

You must submit an electronic copy of your work. Your submission will be electronically checked.

Learner declaration
I certify that the work submitted for this assignment is my own and research sources are fully acknowledged. **Student signature:** **Date:**

Harvard Referencing:

The Harvard Referencing System must be used. The Wikipedia, UKEssays.com or similar websites must **not** be used or referenced in your work.

Introduction

Learning Outcomes:

LO1. Demonstrate the understanding of basic concepts of Big Data, its importance and need in business context.

LO2. Explain the various components of Hadoop and HFDC along with their role in the Big Data ecosystem.

LO3. Summarize the learning on Big Data analytics using Yarn, HDFC and MapReduce

Overview:

This project-based assignment is designed for master's students to demonstrate their knowledge of Big Data concepts and hands-on experience with cloud computing. Students will be required to use **Apache Spark** and **Hadoop** in a cloud environment (AWS EMR, Google Dataproc, or Azure HDInsight, Databricks) to process and analyse a large dataset. The focus of this assignment is on the practical implementation of Big Data technologies and the derivation of meaningful insights from large-scale data.

Assignment Tasks:

1. Problem Definition and Business Context (15% of total marks):

- Identify a real-world business problem or case study where Big Data analytics can be used to drive decision-making.
- Write a brief report (500-800 words) explaining the business context, the need for Big Data, and how large-scale data analytics can provide value in solving the problem.
- Suggest and justify a relevant dataset that will be used for the project (the dataset should be publicly available and of substantial size, >10GB).

2. Cloud Environment Setup and Data Ingestion (25% of total marks):

- Choose a cloud platform (AWS, Google Cloud, or Azure) and set up a **Big Data processing environment** (using EMR, Dataproc, or HDInsight) to run Apache Spark and Hadoop.
- Document the steps taken to configure the cluster, including selecting the appropriate instance types, scaling options, and cost considerations.
- Upload your dataset into **HDFS** and explain the data ingestion process, including handling file formats (e.g., CSV, JSON, Parquet) and ensuring data is properly distributed across nodes.

3. Data Processing with Spark and Hadoop (30% of total marks):

- Implement **two data processing tasks**:

    1. **Hadoop MapReduce** job: Create a basic MapReduce job in Python to process and clean your dataset (e.g., counting word frequencies, detecting anomalies, or aggregating data).
    2. **Apache Spark** job: Use Spark (via Python) to perform advanced data transformations and processing, such as data aggregation, filtering, and exploratory data analysis (EDA).

- Evaluate the performance of both tasks and compare **MapReduce** with **Spark**, considering speed, scalability, and ease of use.

4. Advanced Analytics and Machine Learning (30% of total marks):

- Using **Apache Spark MLlib**, implement a **machine learning algorithm** (e.g., classification, regression, or clustering) on your dataset.

- Provide a detailed description of the **model selection process**, including data preprocessing, feature selection, model training, and evaluation.

- Visualize and explain the **results** of your model, highlighting any business insights derived from the analysis.

Data Source:

1. Amazon Customer Reviews (E-commerce Dataset)

This dataset contains reviews of products on Amazon, providing insights into customer sentiments and product popularity.

- **Size**: >10GB
- **Source**: AWS Public Dataset
- **Use Case**: Sentiment analysis, customer behavior, product trends.
- **Data Link**: https://amazon-reviews-2023.github.io/
- **Cloud Integration**: Easily accessible through AWS S3 and can be processed on AWS EMR or other cloud services.
- **Retail store Data link**: https://github.com/futurexskill/bigdata

Submission Instructions:

- Compile a comprehensive project report or presentation that addresses each task outlined in the preceding section. This report includes:
    - Steps to setup the Hadoop cluster
    - Steps to ingest data into HDFS
    - Source code of the MapReduce job (Mapper and Reducer)
    - Instructions on how to submit the job to YARN
    - Analysis of the results obtained from the MapReduce job
- Ensure that your report is clear, well-organized, and visually appealing
- Prepare a document using the BSBI assignment template available on Canvas.
- User Harvard referencing style for your bibliography.
- Refer to the Essay-Guide available on Canvas for further instructions.
- Submit your assignment electronically by the specified deadline.

GRADING DESCRIPTORS: LEVEL 7

EXPERIMENTATION & INNOVATION								
	FAIL			**PASS**				
Threshold Criteria	**0-29%**	**30-39%**	**40-49%**	**50-59%**	**60-69%**	**70-79%**	**80-89%**	**90-100%**
Deals with complex issues both systematically and creatively demonstrating self-direction and originality in tackling and solving problems	Little to no ability to use techniques to deal with complex issues systematically (including those of ethics and sustainability) and creatively to solve problems and/or make decisions.	Low utilisation of established techniques to deal with complex issues systematically (including those of ethics and sustainability) and creatively to solve problems and/or make decisions, but with limitations in techniques or approach.	Limited research or advanced scholarship to their area of study by using a range of information and established and advanced techniques	Competent understanding of solving problems, through own research or advanced scholarship displaying a comprehensive understanding of established and advanced techniques	Good understanding of solving problems through own research and advanced scholarship critically selecting and displaying a comprehensive understanding of established and advanced techniques.	Very Good problem-solving skills displaying a comprehensive understanding of techniques applicable to their own research or advanced scholarship	Excellent range of extremely well-developed problem-solving displaying an understanding of techniques applicable to their own research or advanced scholarship beyond which is taught.	Exceptional problem-solving skills with sophisticated evaluation and application of a wide range of advanced information and techniques to undertake projects.
Comprehensive understanding of techniques applicable to their own research or advanced scholarship	Little to no understanding of techniques applicable to their own research or advanced scholarship or their limitations and ambiguities.	Low understanding of techniques applicable to their own research or advanced scholarship including their limitations and ambiguities.	Limited understanding of key techniques applicable to their own research or advanced scholarship including their limitations and ambiguities.	Competent understanding of techniques applicable to their own research or advanced scholarship including their limitations and ambiguities	Good understanding of techniques applicable to their own research or advanced scholarship and a some understanding of more specialised techniques.	Very good understanding of techniques applicable to their own research or advanced scholarship and a some understanding of more specialised techniques.	Excellent understanding of techniques applicable to their own research or advanced scholarship and mastery of some more specialised areas.	Exceptional understanding of techniques applicable to their own research or advanced scholarship and mastery of some more specialised areas.

GRADING DESCRIPTORS: LEVEL 7

RESEARCH & ANALYSIS								
	FAIL			**PASS**				
Threshold Criteria	**0-29%**	**30-39%**	**40-49%**	**50-59%**	**60-69%**	**70-79%**	**80-89%**	**90-100%**
Systematic understanding of knowledge, and a critical awareness of current problems and/or new insights, much of which is at, or informed by, the forefront of their academic discipline, field of study or area of professional practice	Little to no knowledge of the subject with limited breadth or depth or deficiencies in major areas or currency.	Low knowledge of the subject lacking coherence, breadth, or detail with only some reference to ideas or arguments at the forefront of any part of the subject.	Limited knowledge to deal with terminology, facts and concepts some of which is informed by the forefront of defined areas of the subject.	Competent knowledge of ideas or arguments at the forefront of any part of the subject sufficient to deal with current issues in the discipline, generally more descriptive than critical or analytical.	Good knowledge of ideas or arguments at the forefront of any part of the subject showing a clear, critical insight into the discipline as whole and current issues/problems.	Very good knowledge of ideas or arguments at the forefront of the subject some of which are significantly beyond what has been taught and show a critical insight into the discipline and current issues/problems.	Excellent knowledge of ideas or arguments at the forefront of the subject many of which are significantly beyond what has been taught and show a critical insight into the discipline and current issues/problems.	Exceptional knowledge of ideas or arguments at the forefront of the subject most of which are significantly beyond what has been taught and show a critical insight into the discipline and current issues/problems.
Conceptual understanding that enables the student to display originality in the application of knowledge	Little to no conceptual understanding or argument and a focus on descriptive explanations which do not comment on arguments of others or alternative views.	Low conceptual understanding and arguments are weak or poorly constructed, and the work does not critically evaluate the arguments of others or consider alternative views.	Limited conceptual understanding and argument construction with critical evaluation of alternative views or comment on advanced scholarship.	Competent conceptual understanding and argument construction with critical evaluation of a range of views and consistent engagement with advanced scholarship.	Good conceptual understanding which critically evaluate and synthesise other views and information with a thoughtful interpretation of advanced scholarship.	Very good conceptual understanding which systematically synthesises a wide range of views with a critical insight into advanced scholarship.	Excellent conceptual understanding which critically apply a wide range of views through a perceptive use of advanced scholarship.	Exceptional conceptual understanding of publishable quality with systematic engagement and usage of advanced scholarship.

GRADING DESCRIPTORS: LEVEL 7

ENGAGING WITH PRACTICE								
	FAIL			**PASS**				
Threshold Criteria	**0-29%**	**30-39%**	**40-49%**	**50-59%**	**60-69%**	**70-79%**	**80-89%**	**90-100%**
Practical understanding of how established techniques of research and enquiry are used to create and interpret knowledge in the discipline	Little to no evidence of background investigation, analysis, research, enquiry, ethical awareness, and/or study.	Low evidence of background investigation, analysis, research, enquiry, ethical awareness, and/or study.	Limited background investigation, analysis, research, enquiry , ethical awareness, and/or study using established techniques, with the ability to extract relevant points.	Competent investigation, analysis, research, enquiry, ethical awareness, and/or study using established techniques accurately, and can critically appraise and use academic sources.	Good background investigation, analysis, research, enquiry, ethical awareness, and/or study using established techniques accurately, and possesses a well-developed ability to critically appraise a wide range of sources.	Very good, independent, extensive and appropriate investigation, analysis, research, enquiry, ethical awareness, and/or study beyond the usual range, and critically evaluates this to advance the work and/or direct arguments.	Excellent independent, extensive and appropriate investigation, analysis, research, enquiry, ethical awareness, and/or study well beyond the usual range, and critically evaluates this to advance the work and/or direct arguments.	Exceptional investigation, analysis, research, enquiry, ethical awareness, and/or study which demonstrates carefully considered depth and breadth and critically synthesises this to advance the work and/or direct arguments.
Originality in the application of knowledge	Little to no technical, creative or artistic skills related to their area of study.	Low technical, creative or artistic skills related to their area of study.	Limited technical, creative or artistic skills required for area of study.	Competent technical, creative or artistic skills required for area of study.	Good technical, creative or artistic skills required for area of study.	Very good range of technical, creative or artistic skills.	Excellent range of technical, creative or artistic skills	Exceptional range of technical, creative or artistic skills
Independently advance your own knowledge and understanding, and to develop new skills to a high level.	Little to no contribution to group activity and/or undertaking further training at a high/advanced level.	Low contribution to group activity and/or undertaking further training at a high/advanced level.	Limited contribution to group activity and/or undertaking further training at a high/advanced level.	Competent contribution to group activity and/or independently undertakes further training at a high/advanced level.	Good contribution to group activity and/or independently undertakes further training at a high/advanced level with an understanding of team roles	Very good contribution to group activity and/or independently undertakes further training at a high/advanced level with an understanding of team roles	Excellent contribution to group activity and/or independently undertakes further training at a high/advanced level with teamwork and leadership	Exceptional contribution to group activity and/or independently undertakes further training at a high/advanced level with teamwork and strong leadership.

GRADING DESCRIPTORS: LEVEL 7

REALISATION & COMMUNICATION								
	FAIL			**PASS**				
Threshold Criteria	**0-29%**	**30-39%**	**40-49%**	**50-59%**	**60-69%**	**70-79%**	**80-89%**	**90-100%**
Communicate information, ideas, problems and solutions to both specialist and non-specialist audiences.	Little to no clarity in the communication of ideas, problems and solutions to audiences.	Low clarity in the communication of ideas, problems and solutions to audiences.	Limited clarity in the communication of ideas, problems and solutions to audiences.	Competent communication of ideas, problems and solutions to audiences.	Good, confident and clear communication of ideas, problems and solutions to audiences in a range of means / media.	Very good, confident and clear communication of ideas, problems and solutions to audiences in a range of means / media.	Excellent communication of ideas, problems and solutions to audiences in a range of means / media.	Exceptional communication of ideas, problems and solutions to audiences in a range of means / media.

GRADING DESCRIPTORS: LEVEL 7

PERSONAL & PROFESSIONAL CONNECTIVITY								
	FAIL			**PASS**				
Threshold Criteria	**0-29%**	**30-39%**	**40-49%**	**50-59%**	**60-69%**	**70-79%**	**80-89%**	**90-100%**
Independently advance your own knowledge and understanding, and develop new skills to a high level.	Little to no contribution to group activity and/or undertaking further training at a high/advanced level.	Low contribution to group activity and/or undertaking further training at a high/advanced level.	Limited contribution to group activity and/or undertaking further training at a high/advanced level.	Competent contribution to group activity and/or independently undertakes further training at a high/advanced level.	Good contribution to group activity and/or independently undertakes further training at a high/advanced level with an understanding of team roles	Very good contribution to group activity and/or independently undertakes further training at a high/advanced level with an understanding of team roles	Excellent contribution to group activity and/or independently undertakes further training at a high/advanced level with teamwork and leadership	Exceptional contribution to group activity and/or independently undertakes further training at a high/advanced level with teamwork and strong leadership.
Qualities and transferable skills necessary for employment requiring: (a) the exercise of initiative, ethical and personal responsibility (b) decision-making in complex and unpredictable contexts	Little to no ability to manage learning and/or exercise initiative, ethical and personal responsibility and/or decision-making in complex and unpredictable situations	Low ability to manage learning and/or exercise initiative, ethical and personal responsibility and/or decision-making in complex and unpredictable situations	Limited ability to manage learning and exercise initiative, ethical and personal responsibility, and decision-making in complex and unpredictable situations	Competent ability to manage learning, and exercise initiative, ethical and personal responsibility, and decision-making in complex and unpredictable situations	Good ability to systematically manage learning, and exercise initiative, ethical and personal responsibility, and decision-making in complex and unpredictable situations	Very good ability to systematically manage learning, and exercise initiative, ethical and personal responsibility, and decision-making in complex and unpredictable situations.	Excellent ability to manage learning on own initiative, and exercise initiative, ethical and personal responsibility, and decision-making in complex and unpredictable situations	Exceptional ability to manage learning on own initiative, and exercise initiative, ethical and personal responsibility, and decision-making in complex and unpredictable situations
	Little to no use of appropriate terminology, limited vocabulary and many errors in spelling, grammar and syntax.	Low use of appropriate terminology, with many errors in spelling, vocabulary and syntax.	Limited expression, style and appropriate vocabulary with errors in spelling, grammar and syntax which affect understanding.	Competent expression, style, and appropriate vocabulary with some errors in spelling, grammar and syntax which do not affect understanding.	Good expression, style and appropriate vocabulary with some errors in spelling, grammar and syntax.	Very good expression, style and appropriate vocabulary with minimal errors in spelling, grammar and syntax.	Excellent expression, style and appropriate vocabulary with minimal errors in spelling, grammar and syntax.	Exceptional expression, style and appropriate vocabulary with no errors in spelling, grammar and syntax.
	Little to no evidence of basic numeracy or digital literacy, hardware and software skills	Low evidence of basic numeracy or digital literacy, hardware and software skills competency.	Limited evidence of numeracy or digital literacy, hardware and software skills competency.	Adequate evidence of numeracy or digital literacy, hardware and software skills competency.	Good evidence of numeracy or digital literacy, hardware and software skills competency.	Very good evidence of numeracy or digital literacy, hardware and software skills	Excellent evidence of numeracy or digital literacy, hardware and software skills competency.	Exceptional evidence of numeracy or digital literacy, hardware and software skills competency.

	competency.					competency.		
	Does not demonstrate achievement of professional competence when assessed against the requirements of a professional, statutory or regulatory body (PSRB).			The student has demonstrated achievement of professional competence when assessed against the requirements of a PSRB.				
	Inaccurate use of terminology with limited vocabulary and many errors in spelling, grammar and syntax. Inaccurate terminology, with many errors in spelling, vocabulary and syntax.			The student has adhered to the appropriate rules and/or conventions set by regulators or the industry.				