Dissertation - Description Report

Liam Thomas

Supervisor: Dr T. R. Khan
Module code: COM6012

October 2025

1 Introduction

The growth of big data and distributed computing has transformed machine learning workflows. Frameworks such as Apache Spark and platforms like Databricks are widely used to power large-scale machine learning pipelines [1, 2, 3]. However, students often find it hard to grasp complex distributed computing concepts such as parallelism, data partitioning, task scheduling and data shuffles. These are the behaviours that set scalable machine learning apart from single machine approaches. This dissertation aims to design and evaluate an interactive visualisation tool that visualises Spark's internal operations as they happen and links them to scalable algorithmic behaviours such as distributed tree splits and neural network updates.

The project draws inspiration from educational creators like 3Blue1Brown and Welch Labs, whose animated explanations have shown how visual and interactive learning can make abstract ideas in mathematics and computing easier to understand [4, 5]. Educational research has also demonstrated that full transparency and interactivity improve comprehension and motivation in technical subjects [6, 7, 8]. By combining these principles with data engineering technology, the project explores how visual storytelling can make scalable machine learning concepts more engaging to learners.

The tool will focus on two main teaching components. The first component will focus on the foundations of Spark, distributed data structures, transformations and cluster execution. While the second will extend these ideas to more advanced topics such as scalable decision trees and neural networks, reflecting the content of the COM6012 module.

Alongside development, this project will evaluate whether the visualisations improve understanding of Spark and scalable ML compared with traditional lectures and labs. The project therefore aims to not only produce a working prototype but also to contribute evidence on how interactive visual methods can support teaching and learning in big data computing.

2 Analysis

2.1 Problem context

The main challenge this dissertation aims to address is the difficulty students face in understanding how distributed machine learning systems, like Apache Spark, actually work and how they are applied. Concepts such as task scheduling, data partitioning, parallelism and data shuffles are central to performance and scalability, yet they are typically taught through abstract code or complicated diagrams. Students often are taught by looking at the results of Spark jobs but not by following the underlying execution flow (how data is distributed, how transformations are scheduled or how communication occurs between nodes. This gap between abstract theory and visualising the processes limits the students' deeper understanding and confidence when working with large scale data frameworks.

2.2 Proposed approach and tech stack

To bridge this gap, this dissertation will focus on building an interactive visualisation tool that demonstrates these distributed processes in real time. The technical design will use a hybrid architecture that keeps the visuals easy to understand while still running real distributed computations. The chosen tech stack combines a Scala backend (type safety and seamless Spark APIs) with Databricks integration for PySpark jobs, which allows real distributed computations to be executed via the Databricks REST API. Results are returned as JSON to a React frontend and rendered with D3.js as interactive visualisations of Spark's internal operations. This combination supports both performance and modularity while aligning with common industry practices. I will start with the COM6012 lab datasets to keep early experiments simple and directly comparable to the taught material. Later, I may add an option for students to import and choose from a wider range of datasets (including their own).

2.3 Design considerations and challenges

A key design decision will be how to explain Spark without oversimplifying its mechanisms and functionality. The tool must accurately represent how Spark executes different jobs while keeping the interface easy

to understand when encountering these ideas for the first time. Potential challenges include managing latency between the backend Spark job execution and the frontend rendering of the visualisations. Another potential challenge would be browser limitations and ensuring the browser does not overload when representing distributed data at a large scale. Another challenge would be that when multiple processes occur simultaneously, visual clarity is maintained.

2.4 Ethics and user study

The project will involve a small-scale user study with university students studying computer science, potentially students enrolled in COM6012. It will provide feedback on whether the visualisations improve understanding and engagement compared with traditional lecture and lab based teaching approaches. To conduct this ethically, this dissertation will require an Ethics Application detailing how the data from the students is collected and stored, and how to ensure student data anonymity. It will also focus on how the feedback survey will be designed and distributed responsibly.

2.5 Feasibility

Overall, this project is feasible given available Databricks resources. The hybrid Scala, Databricks, React and D3.js stack is straightforward yet perfect for demonstrating distributed operations interactively.

2.6 Aim

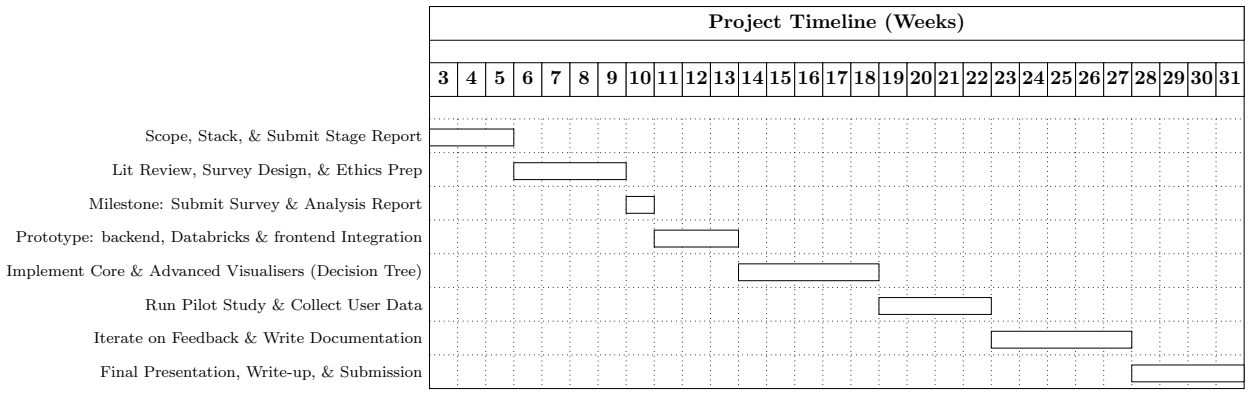
To design and evaluate an interactive visualisation tool that helps students understand how Spark executes scalable machine learning workloads.

2.7 Objectives.

1. Build an interactive visualisation of Spark fundamentals (partitions, stages, shuffles) linked to scalable algorithm behaviour (e.g., distributed decision trees).
2. Get an end-to-end flow running: Scala service → Databricks PySpark → JSON trace → React/D3, with the option to replay or slow down so it works at teaching pace.
3. Start with the COM6012 lab datasets for simplicity. Later consider letting students import and choose other datasets.
4. Run a small, ethics approved pilot to measure learning gains and task accuracy and usability.

3 Plan of Action

3.1 Gantt Chart



3.2 Weekly Plan Breakdown

W3–W5
Finalise the dissertation scope and confirm the tech stack. Read through provided sources and make brief notes. Identify gaps in current teaching tools for Spark. Submit Description Stage Report by end of W3.

W6–W9
Prioritise the literature review for the Survey & Analysis Report. Draft the ethics application.

W10 Submit Survey & Analysis Report, finalise quiz/survey wording and recruit user study participants.

W11–W13
Start prototyping: set up a Scala backend. Use Databricks REST to trigger a PySpark job, fetch JSON, render a basic React/D3 visual.

W14–W18
Implement the visualiser for Spark foundations. Add dataset/job controls and focus on clear step-by-step execution views.

W19–W22
Run a small study (ideally COM6012). Collect consent and feedback. Analyse learning improvement and usability.

W23–W27
Iterate based on results and add documentation notes in the GitHub repo.

W28–W31
Prepare the final presentation, complete the write-up, and submit.

References

[1] Apache Software Foundation, "Apache spark: A unified analytics engine for large-scale data processing." https://spark.apache.org/, 2025. Describes Spark as the most widely used engine for scalable computing, supporting data engineering, data science, and machine learning.

[2] Amazon Web Services, "What is apache spark? – aws introduction to apache spark." https://aws.amazon.com/big-data/what-is-spark/, 2025. AWS highlights Spark's use for large-scale data and machine learning workloads across many industries.

[3] Databricks, "How draftkings uses databricks to power fraud detection machine learning pipelines." https://www.databricks.com/blog/data-intelligence-action-100-data-and-ai-use-cases-databricks-customers#:~:text=DraftKings%20leverages%20Databricks%20to%20power,across%20historical%20and%20live%20data, 2025. Databricks case study on powering large-scale real-time ML pipelines for DraftKings.

[4] 3Blue1Brown, "3blue1brown – youtube channel." https://www.youtube.com/@3blue1brown, 2025.

[5] Welch Labs, "Welch labs – youtube channel." https://www.youtube.com/@WelchLabsVideo, 2025.

[6] L. Morales-Navarro and Y. B. Kafai, "Unpacking approaches to learning and teaching machine learning in k-12 education: Transparency, ethics, and design activities," *arXiv preprint arXiv:2406.03480*, 2024.

[7] I. T. Sanusi, S. S. Oyelere, H. Vartiainen, J. Suhonen, and M. Tukiainen, "A systematic review of teaching and learning machine learning in k-12 education," *Education and Information Technologies*, vol. 28, pp. 5967–5997, 2023.

[8] L. Bothmann, S. Strickroth, G. Casalicchio, D. Rügamer, M. Lindauer, F. Scheipl, and B. Bischl, "Developing open source educational resources for machine learning and data science," *arXiv preprint arXiv:2107.14330*, 2021.