

Project Proposal

06 August 2025 07:06

Name: Omar Qassem

Student Number: 2316700

Course: Computing (Data & Information Systems)

Module: ACCA5033 - Changemakers: Creativity and Value Creation

Assignment: Innovation & Research



Aim and Objectives

09 August 2025 16:57

Aim

To design, implement, and evaluate a generic, cross-language benchmarking framework capable of reliably assessing the performance, scalability, and usability of general-purpose big data processing tools and pipelines — encompassing both batch and stream processing — using real-world datasets.

Objectives

1. **Conduct a literature review** of existing benchmarking methodologies, frameworks, and tools used for big data analysis across different programming languages and processing paradigms.
2. **Define benchmarking criteria and metrics** that cover both quantitative (e.g., execution time, throughput, scalability, fault tolerance, memory usage, cost) and qualitative (e.g., ease of use, maintainability) aspects.
3. **Design a cross-language benchmarking architecture** that can be adapted to multiple programming languages and big data processing frameworks with minimal configuration overhead.
4. **Implement a prototype benchmarking system** and test it across a selection of representative big data frameworks for both batch and streaming workloads (e.g., Apache Spark, Apache Flink, Dask).
5. **Evaluate the framework** by running benchmarks with real-world datasets, analysing results for accuracy, reproducibility, and comparability across different tools and languages.
6. **Assess potential for open-source release** by documenting architecture, usage guidelines, and possible extensions for community adoption.

[1] [2]

Research Philosophy

09 August 2025 17:11

For this project I will employ a pragmatist philosophy. The reason for this is that this philosophy recognises that the value of the benchmarking framework I have proposed lies in its practical applicability across different, varied contexts, rather than in its ability to prove a single, objective truth.

Pragmatism supports the use of both qualitative and quantitative methods to build, test and refine the proposed framework, which will enable it to generate performance insights that are both replicable and adaptable for use with different big data environments.

This approach prioritises outcomes that have real-world effectiveness and applicability, and allows methodological flexibility to suit the nature of the problem presented.

[3][4][5][6]

Search Tools

09 August 2025 18:14

Research databases

As I will be benchmarking big data analysis tools, pipelines and techniques, the research databases I require will need to be appropriately equipped with entries covering computer science, software engineering and data science.

The following is a list of databases, along with their scope & relevance to my project and how I intend to use them:

IEEE Xplore

Scope: Peer reviewed computer science, data engineering & big architecture research.

Use: For finding benchmarking papers & distributed system performance evaluation metrics.

ACM Digital Library

Scope: Core software engineering & computer science conference proceedings and journals.

Use: For finding performance studies, framework comparisons & research surrounding big data/pipeline design science.

ScienceDirect

Scope: Broad science coverage with a very strong computing section.

Use: For finding applied big data analytics case studies and big data frameworks.

Google Scholar

Scope: Broad coverage of most topics. "Grey" literature (literature and research that is not formally published by commercial publishers) included.

Use: Good for casting a very wide net and filtering with CRAAP test.

Search strategies

These are the ways of manipulating databases and search engines to optimise them to their full utility.

Boolean Search

- AND - Narrows searches (e.g. "big data" AND "benchmarking" AND "framework")
- OR (e.g. "big data" OR "benchmarking" OR "framework")
- NOT - excluding unwanted results (e.g. "big data" AND "benchmarking" NOT "framework")
- Quotation marks - Explicitly stating exact words or phrases (e.g. "performance criteria")
- Asterisks (*) - wildcard/truncations for variations of phrases (e.g. bench* would find benchmark, benchmarking, benchmarks etc.)

Research keywords

The specific keywords I can use to research would come primarily from the primary concept of my project. The following is not an exhaustive list of keywords/phrases:

- Big data
- Big data benchmarking
- Data pipeline
- Data pipeline benchmarking
- Data pipeline performance
- Batch processing

- Batch streaming
- Batch processing comparison
- Cross-language benchmarking
- Big data language comparison
- Distributed computing
- Cloud computing
- Data processing frameworks

Concept maps

A concept map visually links main topics with subtopics & related terms. Like a mind map, but with a more structured layout and more concise language.

This helps generate/lay out search combinations & ensures I can systematically explore all relevant areas.

Approach

I aim to take a systematic approach to searching for information that will support me in my project:

1. **Initial broad search** using general keywords in Google Scholar to find papers that will act as a foundation for general ideas
2. Refinement using boolean searches in IEEE Xplore/ACM Digital Library to refine further the general ideas found
3. **Identify themes** by reading extracts of papers to categorise them into their different themes (e.g. "benchmarking methodologies", "Performance characteristics/metrics", "Frameworks")
4. **Snowballing** whereby I can use the references of papers I have found to find even further reading. I can also use the "cited by" links in research databases to find more recent research.

Evaluating sources with the CRAAP test

I will use the CRAAP test to evaluate:

- Currency - Is it current/recent enough for a fast moving field in its infancy like big data? i.e. look for work that is 5 years old or less
- Relevance - Does it relate directly to what I am studying?
- Authority - Is the author reputable (i.e. an academic or expert in the field)?
- Accuracy - Is the work peer-reviewed, with a clear methodology and evidence?
- Purpose - is it research and not marketing material attempting to prove a certain narrative?

[7][8]

Search Strategies

11 August 2025 18:08

Population

For this project, I won't be using a "population" in a traditional sense, but rather data engineering professionals, software engineers and/or researchers - people who are embedded in the big data community and who work with the tools and pipelines that support their work.

I will also be collecting quantitative performance data from tool and qualitative data such as practitioner opinions.

Qualitative Sampling

To capture insights, usability feedback and framework effectiveness, the following data collection methods are appropriate:

- **Interview & focus groups:** Ideal for gathering in depth perspectives and ideas.
- **Field studies:** Observing actual usage of systems or gathering via case studies.

[9]

In order to understand the perceptions & experiences or gather feedback about usability of my framework, I will be using mostly **typical case sampling** using those people such as practitioners from companies that use modern, currently widely adopted tools (e.g. Spark, Databricks, AWS). Using "typical" users of big data frameworks and tools will be beneficial as it will give "realistic, relevant insights".

I will supplement the typical case sampling with **unusual case sampling** in order to capture data from participants with non-typical/extreme use cases.

I can also use **snowball sampling** to find participants in additional, non-standard areas, in order to cover all bases.

Random sampling is not an appropriate population sample to use, as it is hard to come to a consensus with a small pool of experts.

[10]

Quantitative Sampling

For collecting performance data from tools and pipelines, the following data collection methods are appropriate:

- **Benchmarking experiments:** Automating batch and streaming tests across frameworks and languages
- **Surveys:** Capture practitioner opinions, preferences and usage patterns.

I will need to statistically validate the performance metrics across tools used in the big data field. To do this I will use **stratified sampling** which divides the populations into subgroups by themes (e.g. batch vs. stream, languages etc.) and **randomly samples** within each group, for validating datasets and workloads so that I can gather data about a plethora of different processing techniques and languages.

There will be no need to use **systemic sampling** or **cluster sampling** as I will be attempting to determine an overarching, general consensus of performance metrics, so the more data and ideas the better.

[11]

Validity

To ensure I am measuring what I am intending to measure and ensuring it is relevant, I will:

- Align benchmarking metrics with my research objectives
- Pilot test framework
- Use real-world datasets to reflect end-usage

Reliability

To ensure repeated measurements would produce the same results, I will:

- Standardise test environments/configurations
- Automate benchmarking tests
- Perform tests several times and use averages of results

Generalisability

To ensure the results can be used beyond my test cases (in the real world), I will:

- Use a variety of datasets/frameworks
- Include all cases (common and edge)

[12]

Ethical considerations

While the project involves dealing with mostly software, there are human participants involved:

- They must give **informed consent** to participate in my study
- I must keep all data **anonymous and confidential**.
- I must **securely protect data**.
- I must ensure participants do not incur any reputational, professional physical harm
- I must be transparent in my approach so that the results of the study are reputable

[13][14]

Pilot Study

11 August 2025 19:05

A pilot study will help me test the feasibility of my methodology before performing a full study. It will mirror my main study's design, but using a far smaller scale, which keep it manageable while allowing me to use it to refine the project.

To create a pilot study, I will need to perform the following:

1. Define objectives: These will be the same as my full-scale project
2. Design the pilot:
 - a. Select small set of tools to test and small datasets, one batch and one streaming dataset.
 - b. Gather quantitative/qualitative analysis from small group of population
3. Collect data: Run tests in controlled conditions & then gather feedback
4. Analyse test results for issues, obstacles and consistency.
5. Refine: Adjust the configurations of the test as necessary ready for full-scale testing.

[15][16]

Bibliography

17 August 2025 22:57

- [1] D. Phair and A. Shaeffer, "Research Questions, Objectives & Aims (+ Examples)," *Grad Coach*, Jun. 27, 2022. <https://gradcoach.com/research-aims-objectives-questions/>
- [2] OHSU Community Research Coalition HUB, "Establish Your S.M.A.R.T Objectives," [www.ohsu.edu](https://www.ohsu.edu/sites/default/files/2018-12/S-M-A-R-T-101-final.pdf). <https://www.ohsu.edu/sites/default/files/2018-12/S-M-A-R-T-101-final.pdf> (accessed Aug. 12, 2025).
- [3] "(PDF) Mixed Methods Research: A Research Paradigm Whose Time Has Come," *ResearchGate*. https://www.researchgate.net/publication/225083749_Mixed_Methods_Research_A_Research_Paradigm_Whose_Time_Has_Come
- [4] V. Kaushik and C. A. Walsh, "Pragmatism as a research paradigm and its implications for social work research," *Social Sciences*, vol. 8, no. 9, pp. 1–17, 2019, doi: <https://doi.org/10.3390/socsci8090255>.
- [5] D. L. Morgan, "Pragmatism as a Paradigm for Mixed Methods Research," *Sagepub.com*, 2024. <https://methods.sagepub.com/book/mono/integrating-qualitative-and-quantitative-methods-a-pragmatic-approach/chpt/pragmatism-as-paradigm-mixed-methods-research>
- [6] E. Engström, M.-A. Storey, P. Runeson, M. Höst, and M. T. Baldassarre, "How software engineering research aligns with design science: a review," *Empirical Software Engineering*, Apr. 2020, doi: <https://doi.org/10.1007/s10664-020-09818-7>.
- [7] O. Ingram, "How to Apply the CRAAP Test," *Research Prospect*, Sep. 05, 2023. <https://www.researchprospect.com/how-to-apply-the-craap-test/>
- [8] "The theory underlying concept maps and how to construct and use them - Resource," *Better Evaluation*, 2008. <https://www.betterevaluation.org/tools-resources/theory-underlying-concept-maps-how-construct-use-them> (accessed Aug. 17, 2025).
- [9] T. C. Lethbridge, S. E. Sim, and J. Singer, "Studying Software Engineers: Data Collection Techniques for Software Field Studies," *Empirical Software Engineering*, vol. 10, no. 3, pp. 311–341, Jul. 2005, doi: <https://doi.org/10.1007/s10664-005-1290-x>.
- [10] L. Palinkas, S. Horwitz, C. Green, J. Wisdom, N. Duan, and K. Hoagwood, "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research," *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 42, no. 5, pp. 533–544, Sep. 2015, doi: <https://doi.org/10.1007/s10488-013-0528-y>.
- [12] D. Makwana, P. Engineer, A. Dabhi, and H. Chudasama, "Sampling methods in research: A review," *ResearchGate*, 2023. https://www.researchgate.net/publication/371985656_Sampling_Methods_in_Research_A_Review
- [13] D. Siegle, "Research Ethics and Informed Consent | Educational Research Basics by Del Siegle," *Uconn.edu*, Jan. 16, 2023. <https://researchbasics.education.uconn.edu/ethics-and-informed-consent/>
- [14] J. Hecker and N. Kalpokas, "Preserving Privacy and Confidentiality in Research," *ATLAS.ti*, 2024. <https://atlasti.com/guides/qualitative-research-guide-part-1/confidentiality-privacy-research>
- [15] "Carnegie Mellon Mark Kasunic." Accessed: Aug. 17, 2025. [Online]. Available:

https://www.sei.cmu.edu/documents/2984/2004_017_001_22829.pdf

[16] Critical Appraisal Skills Programme, "What is a pilot study?," *CASP - Critical Appraisal Skills Programme*, 2024. <https://casp-uk.net/news/what-is-a-pilot-study>