Individual Plan Master's Thesis

Name: Casper Ove Kristiansson
Email: Casperkr@kth.se

May 15, 2025

1 Project Information

- **Preliminary Title:** From Experiment to Insight: A Comparative Study of Storage Approaches for Large-Scale Synchrotron and Neutron Scattering Data on AWS

- **Examiner (KTH):** Prof. Mats Nordahl

- **Supervisor (KTH):** Prof. Sebastian Dalleiger

- **Industry/External Supervisor:** Dr. Ahmet Bahadır Yıldız (Scatterin AB) — `ahmet@scatterin.com`

- **Keywords:** Synchrotron data; Neutron scattering; Cloud storage; AWS; HDF5; Zarr; Parquet; DynamoDB; Scientific workflows Arrow/Feather; ROOT; Compression;

2 Background & Objective

2.1 Scientific / Societal Context

Modern synchrotron and neutron facilities face issues with data handling where experiments can produce terabytes of data per day, which has overwhelmed traditional storage and analysis methods [1]. During the last decade, numerous strategies have emerged to optimize data storage, management, and access for these large-scale scientific datasets. This thesis will investigate and prototype certain cloud-based solutions (particularly on Amazon Web Services) to handle large volumes of experimental data.

2.2 Interest of Organization (Scatterin AB) & Required Background Knowledge

Scatterin AB currently has a commercial platform that provides materials data analytics. With this platform, the current infrastructure has quite a few drawbacks regarding how it is set up and handles large-scale data sets. I am completing an MSc in Computer Science at KTH with coursework in *Data-Intensive Computing, Data Storage Paradigms, Advanced Algorithms*, and significant experience (2,000+ hours) as a Full-Stack Developer at Scatterin AB, building AWS-based platforms for scientific data.

2.3 High-Level Objective

This thesis intends to *compare* multiple AWS-oriented storage approaches (HDF5, Zarr, Parquet, Arrow/Feather, ROOT, and DynamoDB) using real-world synchrotron/neutron datasets. The goal is to characterize the performance (latency, throughput, compression efficiency), cost, and ease of integration for partial retrieval or queries, basic remote append/delete operations, as well as key operations in beamline and scattering experiments.

3 Research Question & Method

3.1 Research Question

How do six candidate formats HDF5, Zarr, Parquet, Arrow/Feather, ROOT, and DynamoDB, compare in performance, scalability, cost, and compression efficiency when storing large-scale synchrotron and neutron scattering data on AWS, especially for partial retrieval?

3.2 Hypothesis

1. **Columnar Advantage (Parquet):** Columnar storage (Parquet) should excel for queries involving partial-column reads, reducing I/O overhead.

2. **Chunked Object Storage (Zarr):** Zarr's chunk-based design may outperform HDF5 for parallel reads/writes in a purely cloud-native (object store) environment.

3. **HDF5 Familiarity:** HDF5 may remain favorable for certain multidimensional access patterns or legacy HPC-style workflows.

4. **NoSQL (DynamoDB) Trade-offs:** DynamoDB may be beneficial for extremely high concurrency, but less suitable for large-scale partial scans (e.g., retrieving entire columns or slices of data).

5. **Zero-Copy Columnar (Arrow/Feather):** Arrow/Feather will outperform Parquet and other formats on large bulk reads due to its memory-mapped, zero-copy design, yielding lower latency and higher throughput.

6. **ROOT Multidimensional (ROOT):** ROOT's tree-based file structure will offer efficient selective access for nested, multidimensional datasets, matching or exceeding HDF5's performance for complex beamline data slices.

3.3 Objectives

1. Build prototypes for six storage formats (HDF5, Zarr, Parquet, Arrow/Feather, ROOT, DynamoDB) ingesting $\sim$1 TB of two data types, tabular (1 M$\times$20) and dense (x,y) arrays where each tested with and without {zstd, LZ4, gzip}.

2. Produce performance and compression metrics (latency, throughput, compression ratio, CPU overhead).

3. Compute month-long TCO scenarios for hot/warm/cold access tiers

4. Produce a concise decision matrix that highlights trade-offs and recommends when to choose each storage approach.

3.4 Tasks & Challenges

1. **Requirements & Literature Review:** Collect typical data requirements for scattering experiments, referencing the approaches in [1]–[4] and other HPC data management projects. To further improve my knowledge within statistical analysis, I will read a few chapters of the book *All of Statistics: A Concise Course in Statistical Inference*.

2. **Prototype Implementation:**
   - Create a consistent dataset for the six approaches (using AWS S3 or AWS native services).
   - Apply consistent chunking/partitioning logic, so each approach handles partial reads comparably.

3. **Testing and Metrics:**
   - *Dataset Volume:* Terabytes of realistic scattering data.
   - *Performance Metrics:* Read/write latency, throughput, compression ratio, and CPU overhead, storage cost (S3 usage, DynamoDB R/W units, data egress), partial-retrieval overhead.
   - *Scalability Testing:* Evaluate each approach as dataset size, concurrency, and remote append/delete/stream workloads grow.

4. **Analysis:**
   - Compare trade-offs in cost, performance, and complexity. Provide recommendations on when each six approaches is most appropriate for large-scale facility data.

3.5 Method

The methodology framework will be a combination of *experimental benchmarking* with a *design–science research* segment and lastly utilizing *statistical analysis*.

3.5.1 Scientific method framework

- **Variables & hypotheses:** Independent variable: storage format (six approaches). Dependent variables: latency, throughput, compression ratio, cost. For each metric and each pair of formats, whether their mean values are equal (no difference) or not, using a 5 % significance level.

- **Design & data:** $n = 50$ randomized runs per format on an EC2 instance, both on a RAM-disk (no swapping) and via AWS S3.

- **Analysis & decision rule:** Shapiro–Wilk for normality; if satisfied, one-way ANOVA + Tukey HSD, otherwise Kruskal–Wallis + Dunn (Holm correction). Report effect sizes; reject when lower than a 5 % significance level.

3.5.2 Experimental benchmarking

Six functionally equivalent storage backends, HDF5, Zarr, Parquet, Arrow/Feather, ROOT, and DynamoDB, will ingest the same ∼1 TB synchrotron/neutron dataset. Each experiment is repeated with and without the three selected compression schemes.

3.5.3 Design–science research cycle & Statistical analysis

Will adopt an iterative build–evaluate–refine process: each prototype is implemented, measured against our performance and reliability targets, and then adjusted based on the results. Insights from one iteration directly inform the design of the next, ensuring continuous improvement. Will use one-way ANOVA to compare mean values of our key metrics (read/write latency, throughput, cost) across the four storage approaches. If the data violate normality or equal-variance assumptions and apply a Kruskal–Wallis test instead.

3.6 Ethics & Sustainability

One of the big things is data privacy. I need to ensure that the data that will be used in the experiments is something that can be published. Regarding sustainability, I will need to discuss energy usage when utilizing cloud environments and compare it to on-premises solutions, and how it affects the carbon footprint.

3.7 Limitations & Risks & Mitigations

This thesis will only focus on AWS, meaning that no other cloud providers will be used to compare. I also don't compare with on-premises solutions using different storage solutions, such as tape. A few risk scenarios are AWS budget over usage, which can be mitigated using cost alerts and ensuring I can utilize AWS credits. Insufficient data volumes can be mitigated by using synthetic data generation as an alternative method. Another big risk is deviation from the schedule, which can be mitigated by always referring back to the plan and having meetings with both the host company and the supervisor to discuss solutions.

4 Evaluation & News Value

4.1 Evaluation Metrics

Will assess each storage approach using concise, single-threaded benchmarks. Read and write latency will be measured by the median and 95% percentile. Compression efficiency combines the compression ratio with CPU time per GiB for zstd, LZ4, and gzip. Cost analysis models a one-month total cost of ownership for a 1 TB dataset in AWS hot, warm, and cold tiers. Finally, partial retrieval overhead is quantified by measuring end-to-end access times for small slices of tabular and dense array data.

4.2 Scientific Relevance & Innovation / News Value

Scientific Facilities: Synchrotrons, neutron sources, and other big-science labs need robust, future-proof strategies. **Data Management Community:** Expands best practices on HPC-cloud integration, partial reads, and cost-optimized big-data solutions, especially in the synchrotron/neutron community.

5 Pre-Study

Collect typical data requirements for scattering experiments, referencing the approaches in [1]–[4] and other HPC data management projects. Will survey around 30 sources, which will be grouped into five different areas: Synchrotron and neutron big-data architectures, Cloud-based beamline portals and remote workflows, HPC I/O libraries and object-storage formats, Cost-aware cloud storage and economics, and lastly, Benchmarking and statistical foundations.

6 Conditions & Schedule

6.1 Resources Needed and External Supervisor Involvement

Three areas: AWS credits (which can be utilized across AWS products), Software and tools (Python, PyArrow, h5py, zarr-python, AWS SDK), Real beamline datasets (provided by Scatterin AB) + synthetic generator. Work carried out primarily at Scatterin AB's KTH campus office, with daily supervision opportunities.

6.2 Timeline

Task	Start	End
Pre		
Kickoff & detailed schedule	2025-05-05	2025-05-11
Literature survey & related work	2025-05-05	2025-05-25
Define thesis scope & research questions	2025-05-12	2025-05-18
Experimental plan & metric design	2025-05-19	2025-05-25
Dataset acquisition & preparation	2025-05-26	2025-06-01
Draft Introduction & Background	2025-05-26	2025-06-08
Consolidate pre-study package for supervisor	2025-06-02	2025-06-08
Main		
Prototype implementation (six different approaches)	2025-06-16	2025-06-29
Data ingestion & validation	2025-06-23	2025-06-29
Benchmarking framework & Methods draft	2025-06-30	2025-07-06
Self-driven experiments & drafting (no external feedback)	2025-07-01	2025-08-31
Wrap up		
Cost modelling & TCO analysis	2025-07-28	2025-08-17
Statistical analysis & visualisation	2025-08-04	2025-08-10
Draft Results & Decision matrix	2025-08-11	2025-08-17
Draft Discussion & Conclusions	2025-09-01	2025-09-07
Self-review & internal revision	2025-09-01	2025-09-07
Final formatting & proofreading	2025-09-08	2025-09-21

References

[1] Y. Wang, N. Li, Z. Lu, *et al.*, "Synchrotron Big Data Science," *Small*, vol. 14, no. 46, e1802291, 2018. DOI: 10.1002/smll.201802291. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30222245/.

[2] G. R. Meyer, D. Thompson, J. M. Guss, *et al.*, "Operation of the Australian Store.Synchrotron for macromolecular crystallography," *Journal of Applied Crystallography*, vol. 47, no. 4, pp. 1548–1560, 2014. DOI: 10.1107/S1600576714012410. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4187999/.

[3] T. Moriyama, T. Watanabe, H. Tominaga, T. Shoji, M. Ito, and K. Tamura, "Public Cloud-based Remote Access Infrastructure for Neutron Scattering Experiments at MLF, J-PARC," in *Proceedings of the 17th International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS)*, 2019. [Online]. Available: https://accelconf.web.cern.ch/icalepcs2019/papers/tubpl04.pdf.

[4] W. F. Godoy, N. Podhorszki, D. Pugmire, S. Klasky, *et al.*, "Efficient Data Management in Neutron Scattering Workflows," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021. arXiv: 2101.02591. [Online]. Available: https://arxiv.org/pdf/2101.02591.