MSc Degree Project Proposal
Computer Science

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

January 25, 2025

1 Thesis Title

From Experiment to Insight: Real-Time Data Analysis and Streamlined Management of Large-Scale Scientific Datasets for Synchrotron and Neutron Scattering

2 Background

This project is based on scientific data management, where the project will address the challenges of handling large-scale datasets from synchrotron and neutron scattering experiments. The goal is to improve the bridge between research and practical development to optimize data storage, processing, and visualization.

2.1 Research Area

The project will focus on scientific data management in terms of optimizing storage, processing, and visualization of large-scale experimental datasets generated by synchrotron and neutron scattering experiments.

2.2 Current Research and Development

- **Emerging Need:** Advances in experimental technologies especially in the field of synchrotron and neutron scattering have shown the exponentially increased volume and complexity of raw scientific data.
- **Cloud Computing and Scalability:** The project will utilize cloud computing on the platform AWS to explore scalable data management and processing solutions.
- **Visualization:** Efficient data visualization is growing in demand for real-time extraction in data-driven areas such as scientific research.

2.3 Interest and Relevance of the Project

- **Scientific Community**: This project provides researchers with efficient methods to manage and analyze large datasets efficiently.
- **Industry**: Building innovative cloud computing solutions for data storage and processing.
- **To the Hosting Organization Scatterin AB**: The project supports the development of advanced tools for data-driven research to improve the analysis of large-scale scientific datasets.

3 Research Question

How can we design and implement an efficient, cost-effective, and scalable data pipeline — from raw data storage to real-time analysis and streamlined visualization — that meets the unique demands of large-scale synchrotron and neutron scattering experiments?

4 Hypothesis

Optimizing data storage, processing, and visualization workflows for large-scale scientific datasets will result in improvements in cost-efficiency and scalability.

1. **Storage Optimization**: Implementation of a tailored storage strategy to reduce cost and improve data retrieval that fits the specific data loading requirements.
2. **Data Preparation**: Data transformation pipeline to enable faster analysis workflows.
3. **Processed Data Storage**: Precomputing analysis results in optimized formats to minimize both storage requirements but also enable downstream tools.
4. **Visualization Efficiency**: Data handling methods to reduce data transferring sent to the visualization tool.

5 Research Method

Research methods for this project will follow a design-science and empirical evaluation approach:

1. **Requirements Gathering & Literature Review**: Understanding the domain and identifying the specific data handling needs and constraints for synchrotron and neutron scattering experiments. Review existing solutions (data lakes, HPC storage, serverless pipelines) and other relevant industrial literature to understand the best practices in this field.
2. **Prototype Design & Implementation**: Propose and formulate a data-pipeline architecture and storage strategies. Will utilize iterative development to build a minimal viable prototype and then integrate the experimental workflows allowing for ETL (extraction, transformation, and loading).
3. **Experimental Setup & Data Collection**: The project will use representative datasets from synchrotron/neutron experiments to mimic real-world scenarios. Collect performance data (qualitative data on throughput, latency, resource utilization, cost, scalability). During the project, qualitative data will be collected on challenges such as development efforts, and operational overhead.
4. **Performance Evaluation & Analysis**: The project will utilize comparative testing where evaluation and comparison of different strategies. It will also consist of model and measuring the total cost of ownership including storage cost, data transfer fees, and computing costs for transformations. The assessment will also include scalability and accessibility assessing how well to handle the increase of users/volume of data.

6 Background of the Student

I am currently pursuing a Master of Science in Computer Science at KTH Royal Institute of Technology, building on my Bachelor's degree in Computer Engineering from the same institution. Through coursework in **Data-Intensive Computing, Data Storage Paradigms, Advanced Algorithms,** and related subjects, I have developed a solid theoretical foundation for large-scale data management and analysis.

On the practical side, I have worked for 2,000 hours as a consultant with hands-on experience as a **Full Stack Developer at Scatterin AB** (the organization where the degree project will be held). Over the past two years, I have been responsible for designing and maintaining a fully cloud-based platform for analyzing scientific material data generated by neutron and synchrotron experiments. My work included:

- **Building and Orchestrating AWS Services**: Managing a multi-stage cloud infrastructure with services such as S3, DynamoDB, AWS Lambda, and more, using Terraform for infrastructure as code.
- **Developing Scalable Data Pipelines**: Creating 100+ Lambda endpoints to handle large-scale data processing and real-time experiment analysis.

- **Designing User-Facing Tools**: Implementing a custom frontend to manage hundreds of gigabytes of experiment data, streamlining data interaction for end users.
- **Cloud Optimization and Collaboration**: Working closely with data scientists to deploy algorithms efficiently in the cloud, improving both performance and cost-effectiveness.

Additionally, my entrepreneurial background—co-founding multiple startups—has led to building skills in **scalable solution design** and **rapid product iteration**, which are important for designing and building high-volume data pipelines. My Bachelor's thesis on **Cloud Computing Pricing and Deployment Efforts** further solidifies my understanding of cloud cost models and implementation strategies, positioning me well to investigate the technical and cost-related aspects of large-scale data storage and analysis.

Given this combination of **academic training** and **significant real-world experience** in cloud infrastructures and large-scale data analysis, I feel well-prepared to conduct this degree project on optimizing data pipelines and storage strategies for synchrotron and neutron scattering experiments.

7 Supervision at the Company/External Organization

This degree project will be conducted at Scatterin AB, a company focused on cloud-based data analysis solutions for synchrotron and neutron experiments. I have already been working at Scatterin AB for the past two years (over 2,000 hours) as a Full Stack Developer, which has given me in-depth familiarity with their infrastructure, data workflows, and overall research focus.

7.1 Supervisor

Ahmet Bahadır Yıldız (PhD, KTH Royal Institute of Technology, within SwedNess: Swedish graduate school on neutron scattering; Metallurgical Engineer from RWTH Aachen). Yıldız has extensive experience working on materials and process optimization using experiments at large-scale facilities. He will provide daily supervision, guiding the project's technical and research directions.

7.2 Additional Company Expertise

Peter Hedström, Co-founder and CSO at Scatterin AB, and Professor in Materials Science at KTH Royal Institute of Technology, has conducted research at ten different neutron and synchrotron X-ray facilities since 2002. While not the direct supervisor for this thesis, his expertise may be consulted for additional insights into large-scale experimental data handling.

7.3 Location

The work will be carried out primarily on the premises of Scatterin AB. Their office is located on the KTH campus at "A Working Lab", providing daily opportunities for in-person supervision and seamless collaboration with the Scatterin AB team.

8 Resources

Scatterin AB already has a fully functioning cloud-based platform designed for analysis of neutron and synchrotron experiment data which I developed from the ground up. The company which is research-focused has in-house data analysis experts who specialize in scientific datasets in large-scale facilities (synchrotron and neutron). The company will provide access to historical and ongoing experimental data from industrial and academic collaborations, providing real-world test cases for new optimizations.

9 Eligibility

I have verified that I'm eligible to start my degree project that I fulfill the basic requirements of starting the project and that all relevant courses are already completed for the project.

10 Study Planning

Currently, the only remaining course I need to complete for my Master of Science in Computer Science is the Program Integrating Course in Computer Science (DD2300), which will run in parallel with the degree project. Additionally, I was enrolled in Scalable Software Development with Functional Programming (DD2489) during the most recent period, which ended one week ago. I have completed a substantial portion of the course requirements and am currently in discussion with the instructor to finalize the remaining components.