Dolosse - A Modern, Scalable, and Extensible Data Acquisition and Management System

VII SPD Collaboration Meeting
23 May 2024

Thabang Mokoena, Casey Callaghan, Shane Carelse, Avesh Sook, Katlego Machethe, Olebogeng Khake
Dr Stan Paulauskas (Project Science)
I Software Engineering Division I R&DTS Department I NRF iThemba Laboratories for Accelerator Based Sciences

Dolosse A Modern, Scalable, and Extensible Data Acquisition and Management System

Agenda

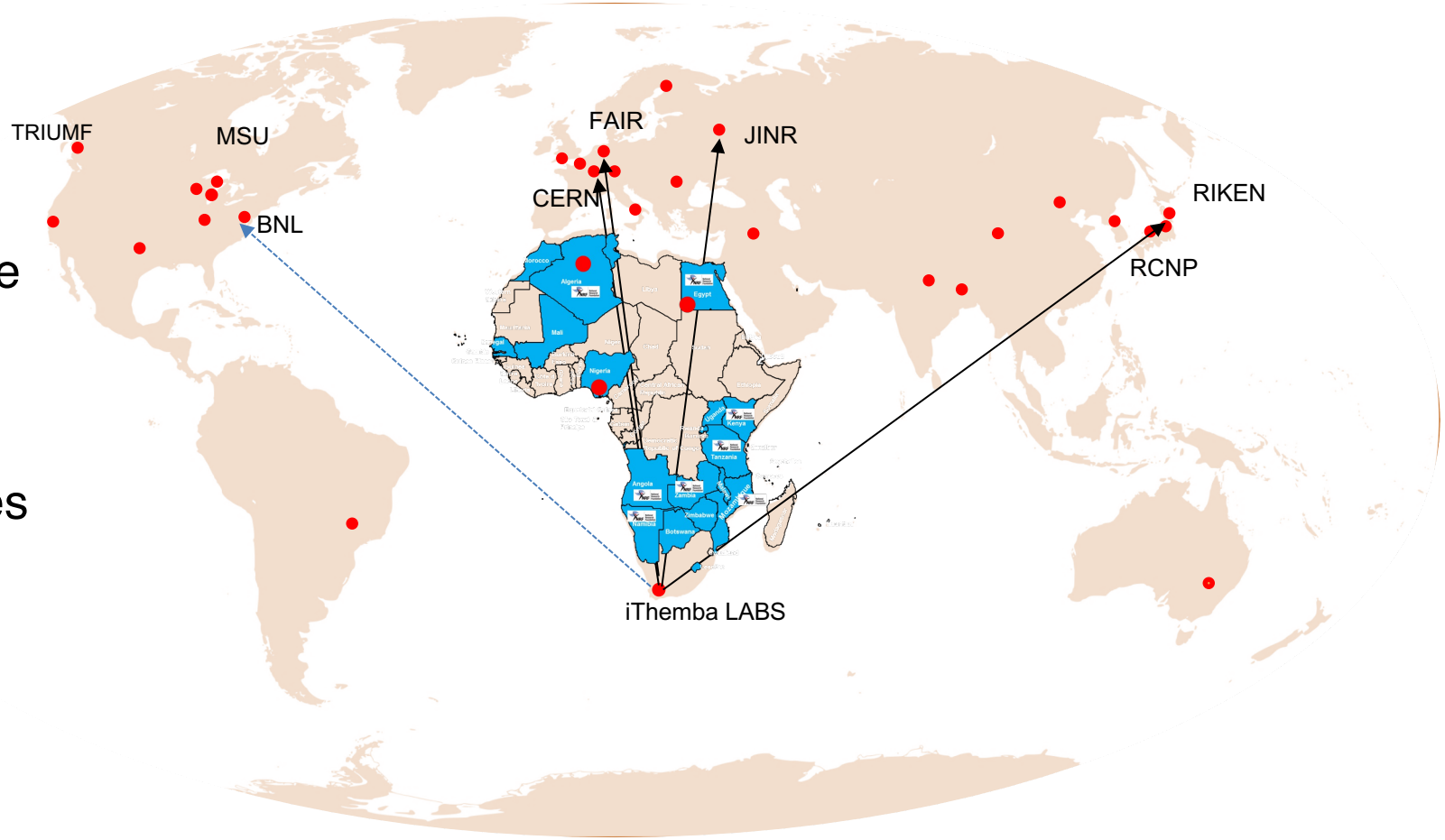
- About iThemba LABS
- Motivation
- Background
- Trends
- Objective
- Dolosse Architecture
- Messaging System
- Storage / Visualization
- SPD Collaboration Plan
- Conclusion

NRF iThemba LABS

- The largest facility of the kind in the Southern Hemisphere and one of the largest in the world

- The African gateway to International Large Scale Research Infrastructures

- Two sites: Cape Town Main site and Gauteng (Wits University premises) satellite campus
- Staff Complement of ~ 290
- 52 Scientists / Researchers

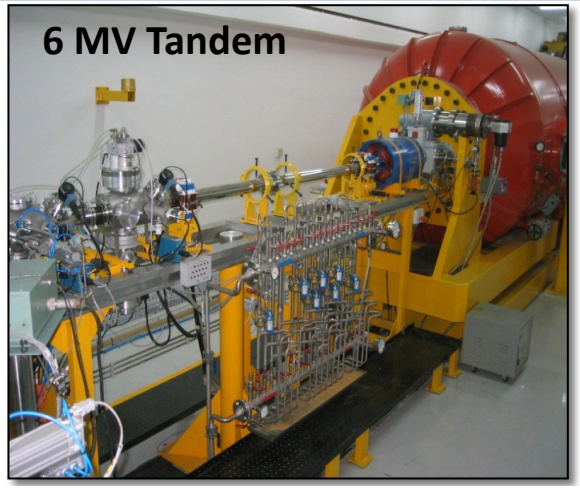


iThemba LABS: Laboratories for Accelerator Based Science

Research Focus

- Fundamental studies of nuclear phenomena;
- Applications of ion beams (IBA) and associated techniques in materials and nanoscience research;
- Accelerator mass spectrometry (AMS);
- Nuclear Medicine – Radionuclides
- Radiation Biophysics - Radiobiology

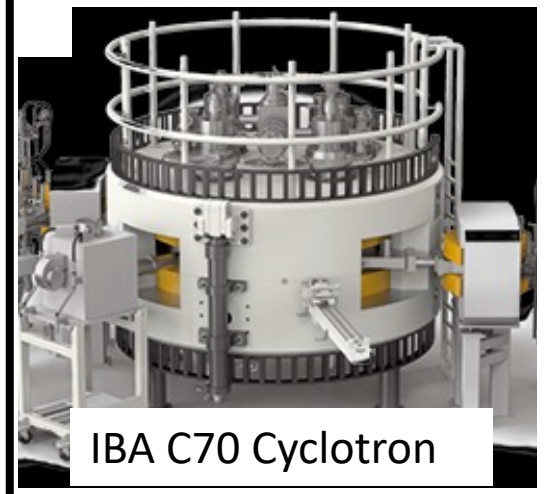
Subatomic Physics and Applications, Nuclear Medicine



Materials Research and Accelerator Mass Spectrometry (AMS)

Materials Research & Nanoscience

Radioisotope Production



South African Isotope Facility

Motivation: Existing Challenges

Reliance on outdated software and hardware

- The continued use of obsolete technology restricts advancements and leads to increased maintenance.

Challenges: inefficiency, lack of flexibility, and support

- These limitations hinder adaptability and progress, leading to prolonged problem-solving times and reduced support options

Impact on experimental outcomes and data integrity

- We are aware that the reliance on aged systems can compromise the accuracy and reliability of research findings and data quality.

Typical Data Flow

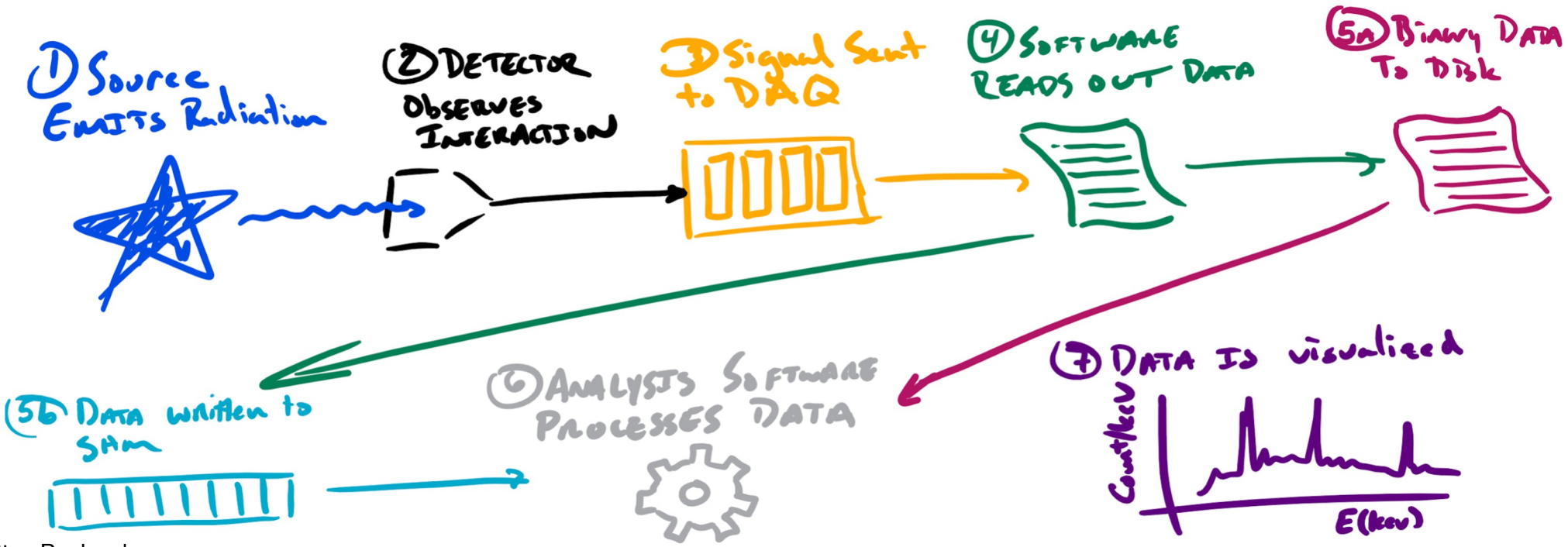


Image source: Dr Stan Paulauskas

- Solid, well-established workflow.
- Feedback loop is short for small experiments.
- Visualization is simple since it's an aggregate
- Can replay data from disk

- Serial processing of data can be slow.
- Correlation across systems is difficult.
- Not easily parallelizable due to the nature of software.
- Need to unpack data every time we analyze.

Trends from Commercial Sectors

Physics

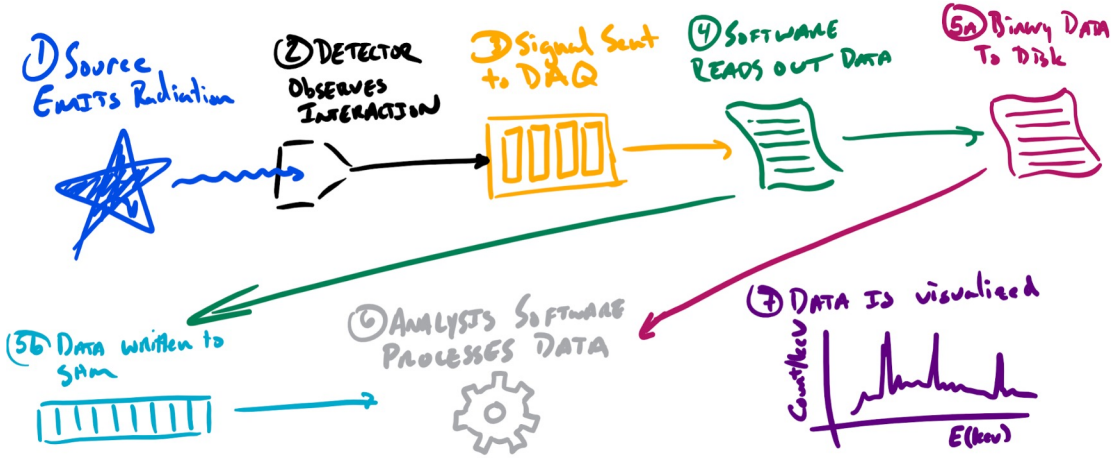
- Processes streaming data from detectors
- Analyzes data in real time
- Needs high-fidelity data storage
- Needs to analyze data from disk

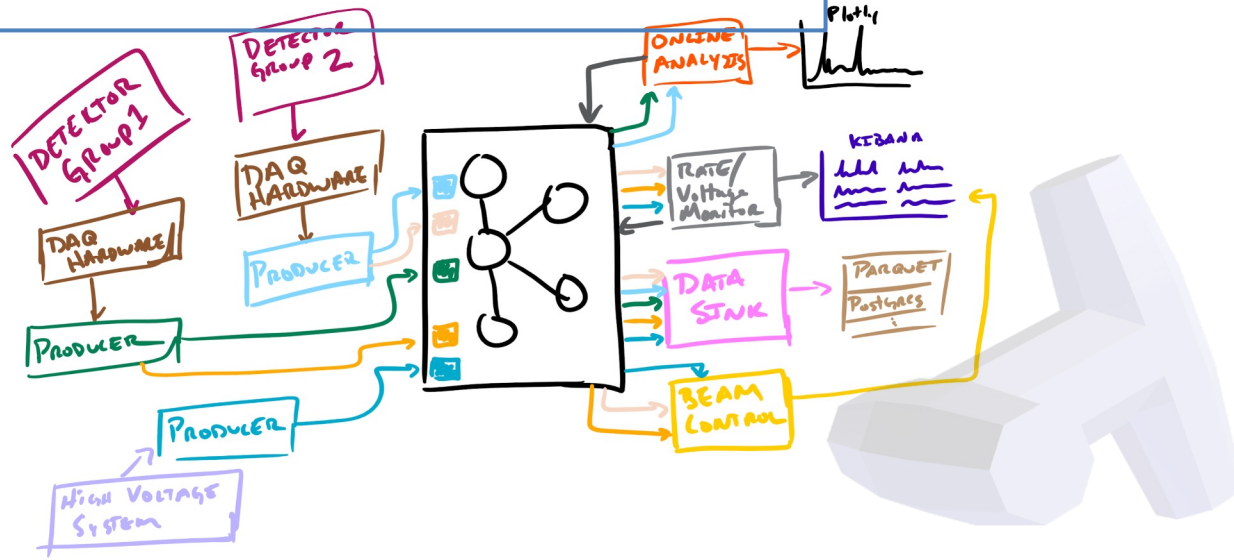
Industry

- Processes streaming data from event streams
- Analyzes data in real time for business decisions
- Needs high-fidelity data storage.
- Analyzes CSV, database entries, etc.

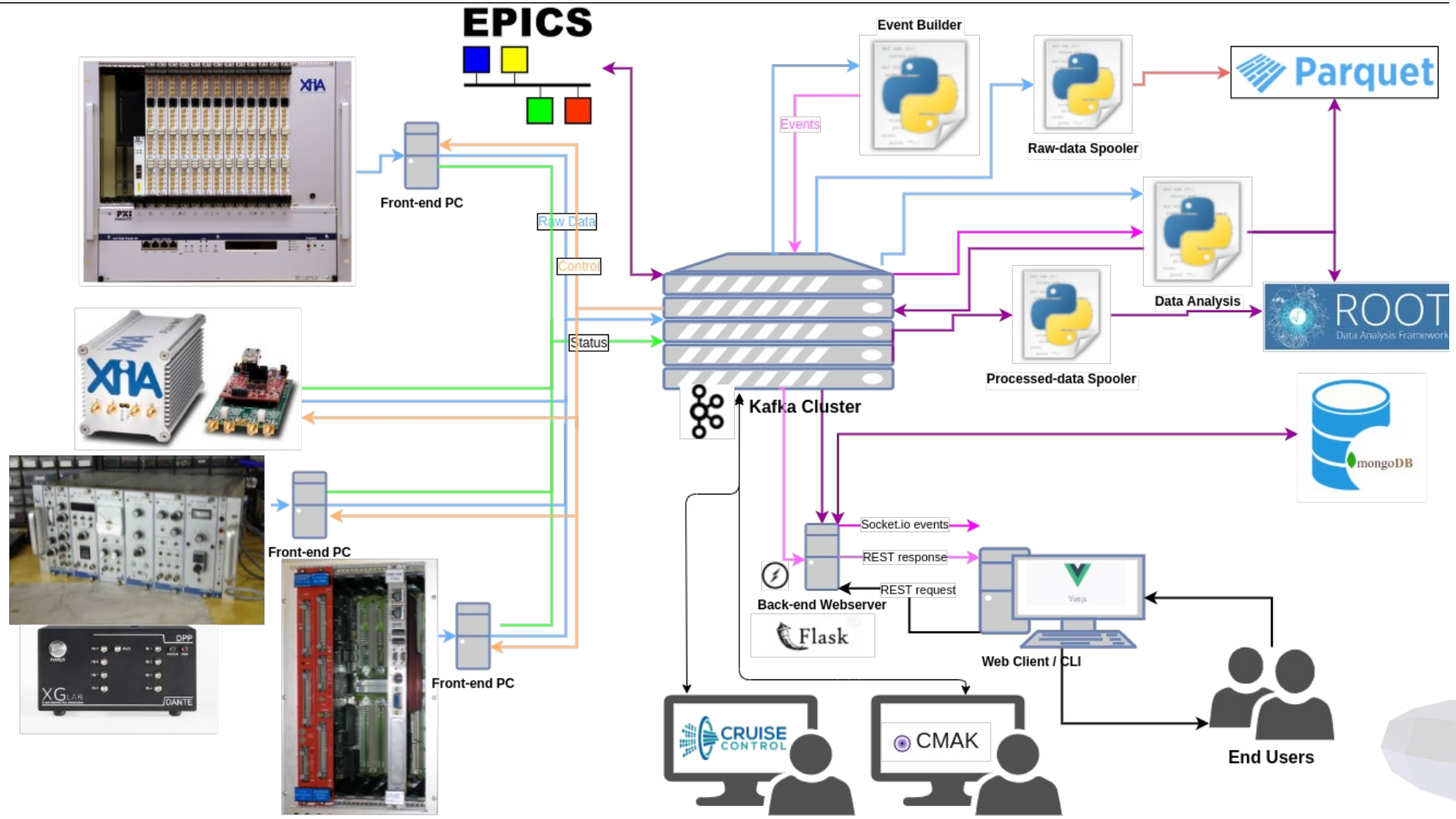
Objective of the Dolosse Project

- Modernize data acquisition and analysis workflows
- Update and streamline processes to handle data more effectively and accurately
- Take advantage of open, supported software solutions
- Platform agnostic remote access to the system
- Data storage to replace binary formats





Dolosse High Level Architecture

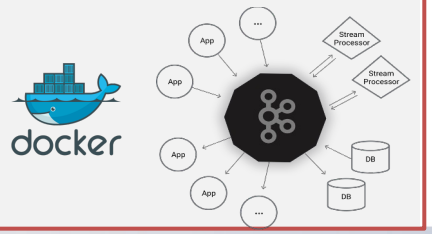


The Approach

Data Acquisition

Microservices

- Kafka is a messaging framework that allows us to manage communication between all of our different data sources. It allows for a multi-producer, multi-consumer model. We can collect and analyze data in real-time.
- Fault-tolerant, replicated data streams
  - Input : 1 M msg / sec I Output: 2 M msg / sec
- Huge amount of community support
- Data retention capability
- Docker for Containerisation



Analysis

Python

- The de facto data-analysis programming language. Incredibly simple to perform advanced analysis of data (e.g. ML)

Web Based Visualisation

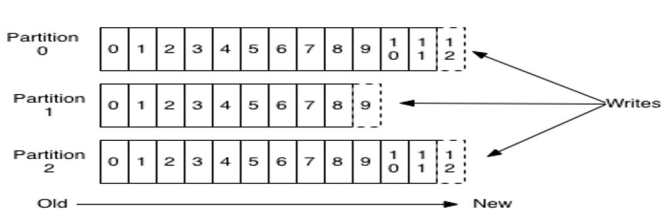
Plotly I Vue.js

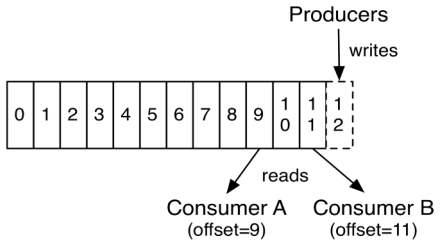
- Plotly excels in creating interactive plots, which greatly enhance the interpretability of our data. While it is open source and offers a free tier, there are additional features available with its paid version that we are currently not utilizing.
- GUI PSI MIDAS Inspired.

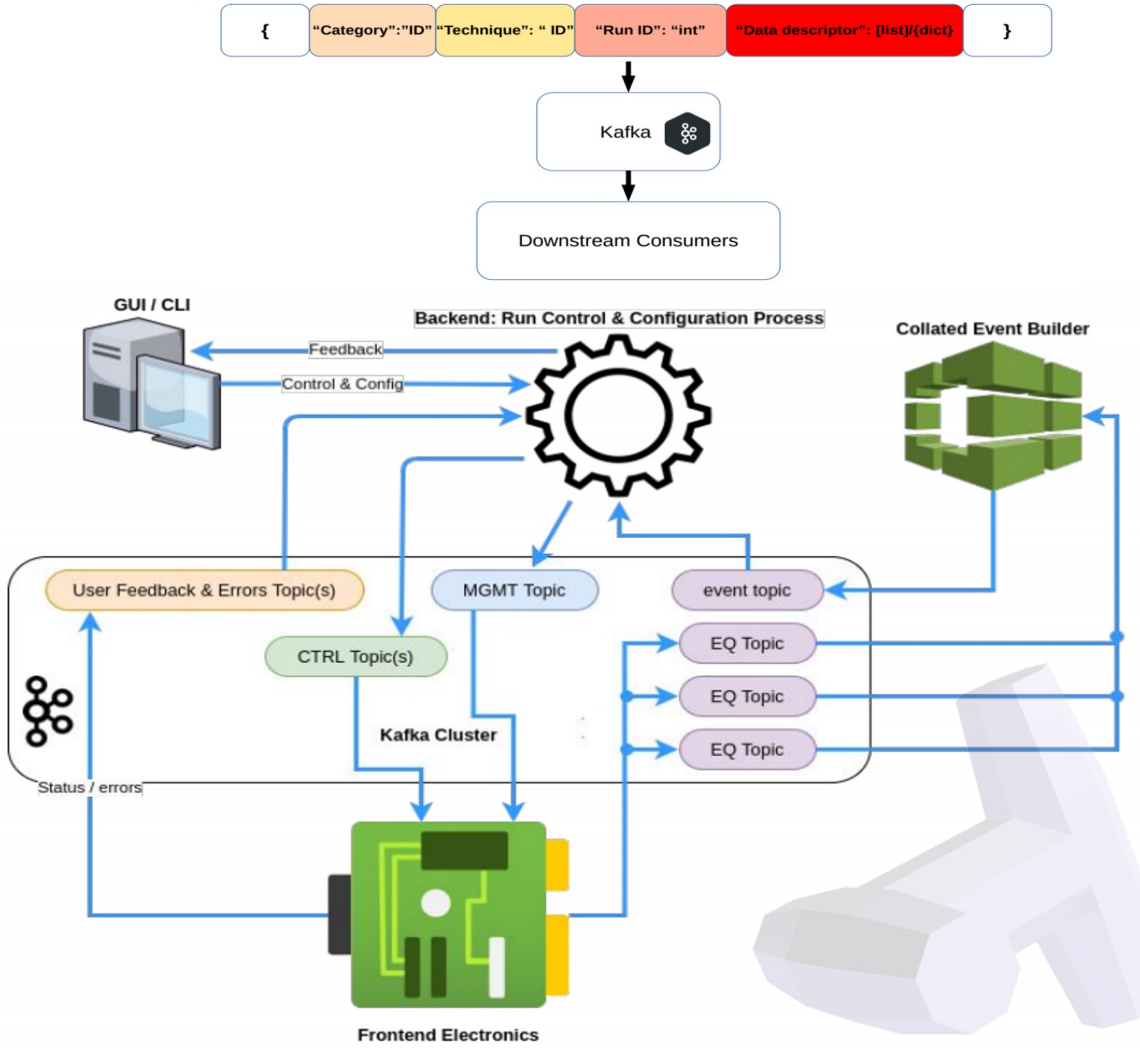
Dolosse Messaging Design

Topics:

- Each topic has partitions that producers write into.
- New messages are always written to the end of the topic.
- Producing and consuming from a topic do not interfere with each other.
- Consumers do not need to be in sync, or start at the beginning of the stream.
- The message payloads can be any binary data.

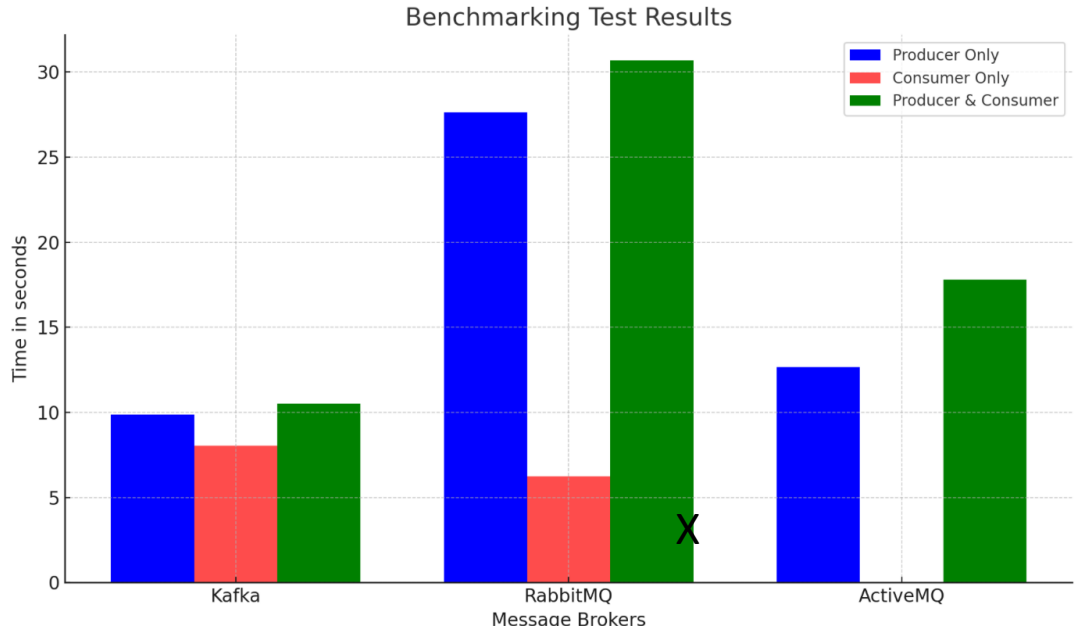






Message Broker Benchmarking

- **Apache Kafka** was the most efficient both in standalone and combined producer-consumer scenarios, balancing high throughput with low latency.
  - Producer Only: Took approximately 9.88 seconds to produce 100,000 messages, showcasing its high efficiency in message production.
  - Consumer Only: Consumed 100,000 messages in about 8.04 seconds.
  - Producer & Consumer: Operating simultaneously, Kafka managed to handle both producing and consuming messages within 10.51 seconds, demonstrating excellent overall throughput and efficiency.



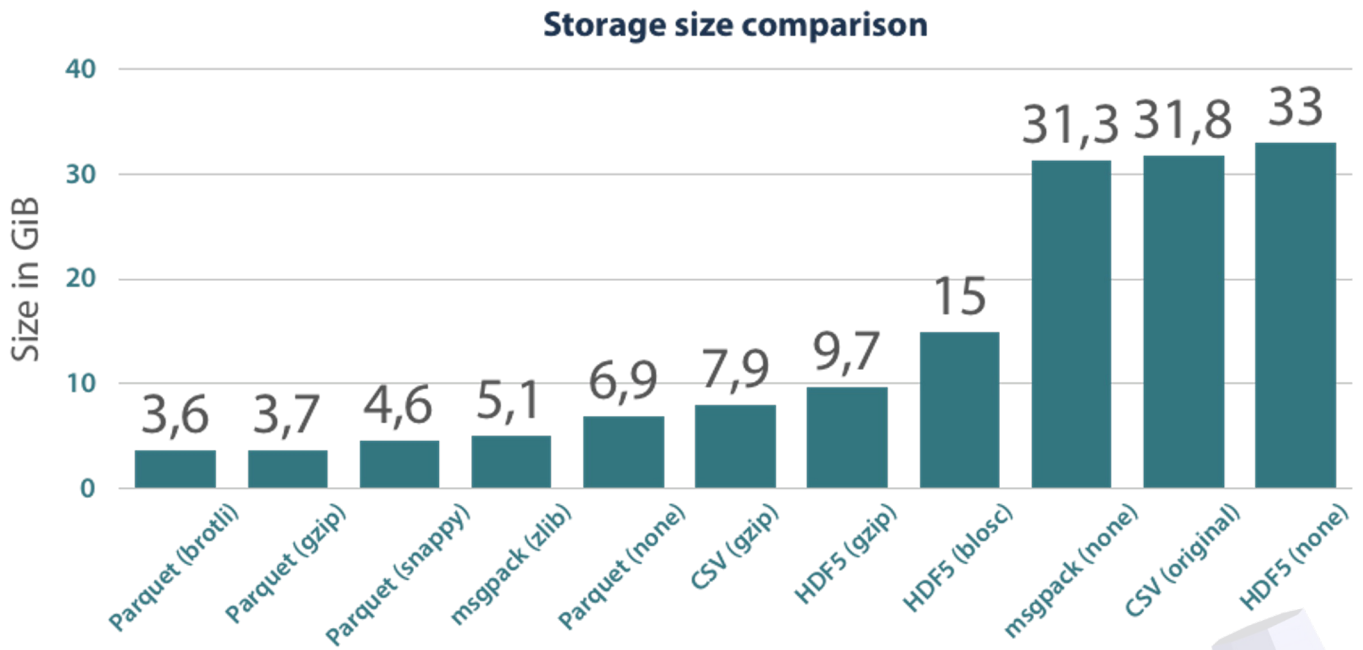
- **RabbitMQ** shows its strength particularly in message consumption speed but falls behind in production speed and combined operation efficiency.

- **ActiveMQ** offers a middle ground in production speed but lacks data on consumption speed for a comprehensive comparison. Its combined operation performance is noticeably slower than Kafka but faster than RabbitMQ.

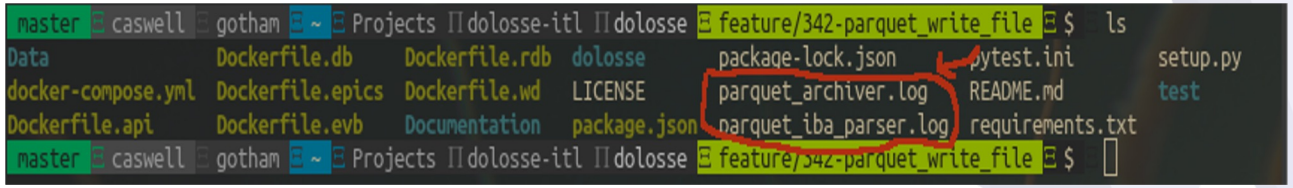
Data Storage Format & Logs

Apache Parquet File Format:

- Apache Parquet files are incredibly efficient storage formats.
- System is designed to actively create logs of the running session, system health, and error information.

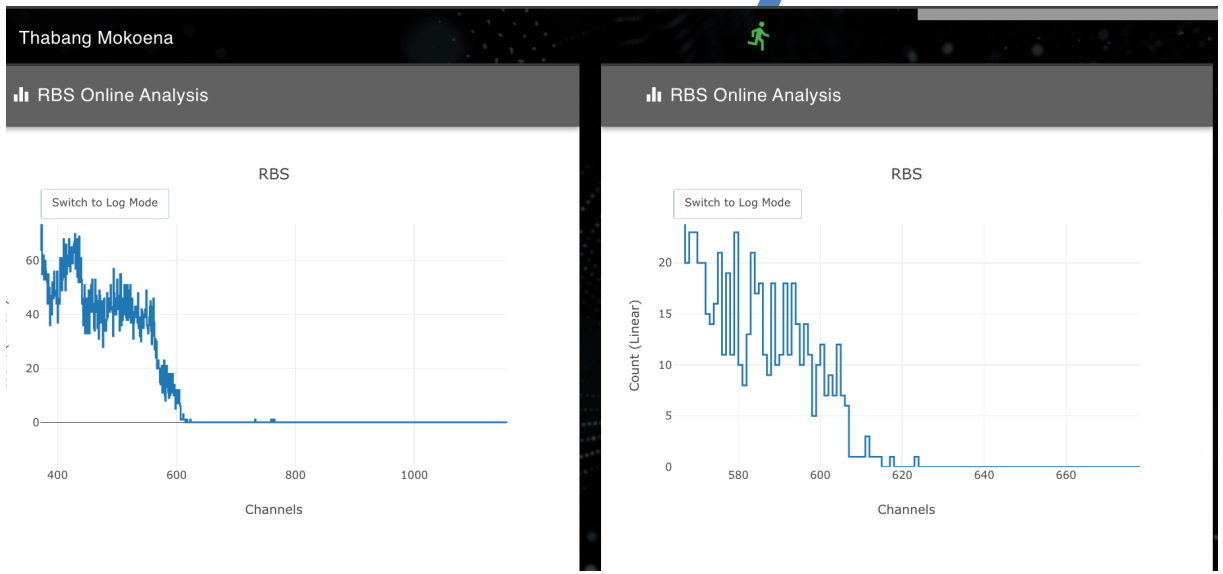


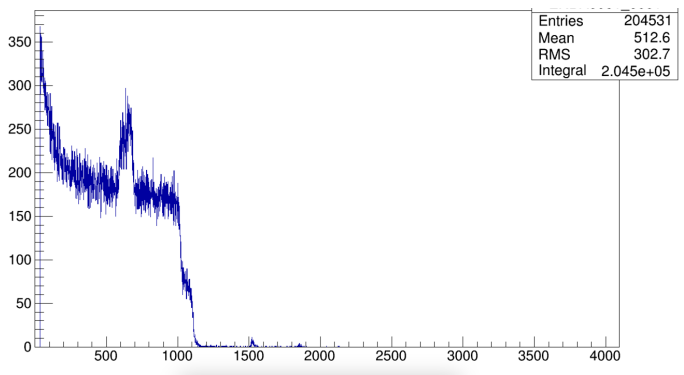
From https://tech.jda.com/efficient-dataframe-storage-with-apache-parquet/

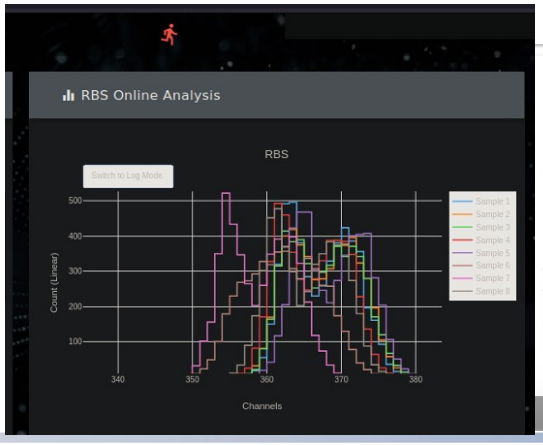


Visualization with Plotly

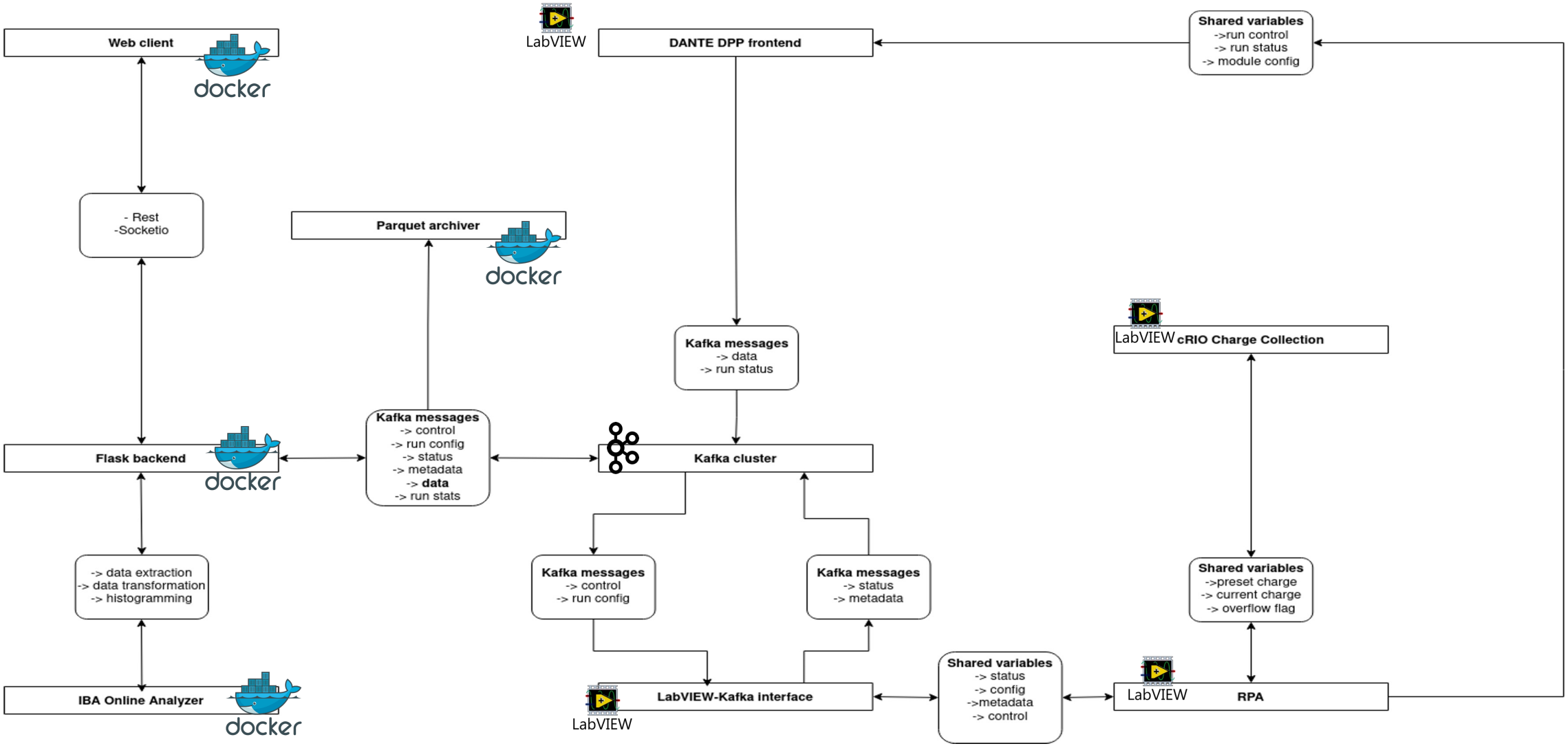
- Although the “best” framework for visualization is really going to be application dependent .Current implementation uses Plotly for live data visualisation.
- Graphically user interface Vue.js framework
- Plotly is great in creating interactive plots
- ROOT integration on next iteration.



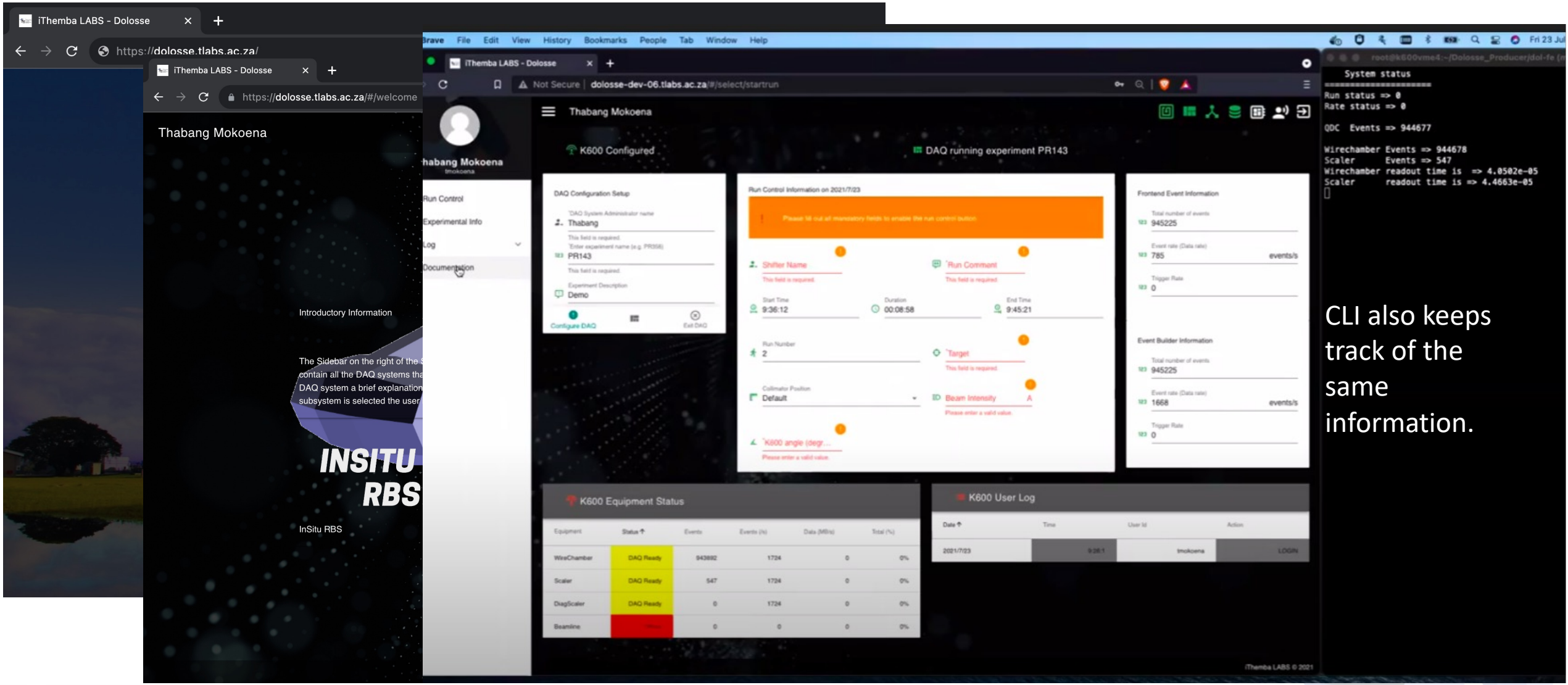




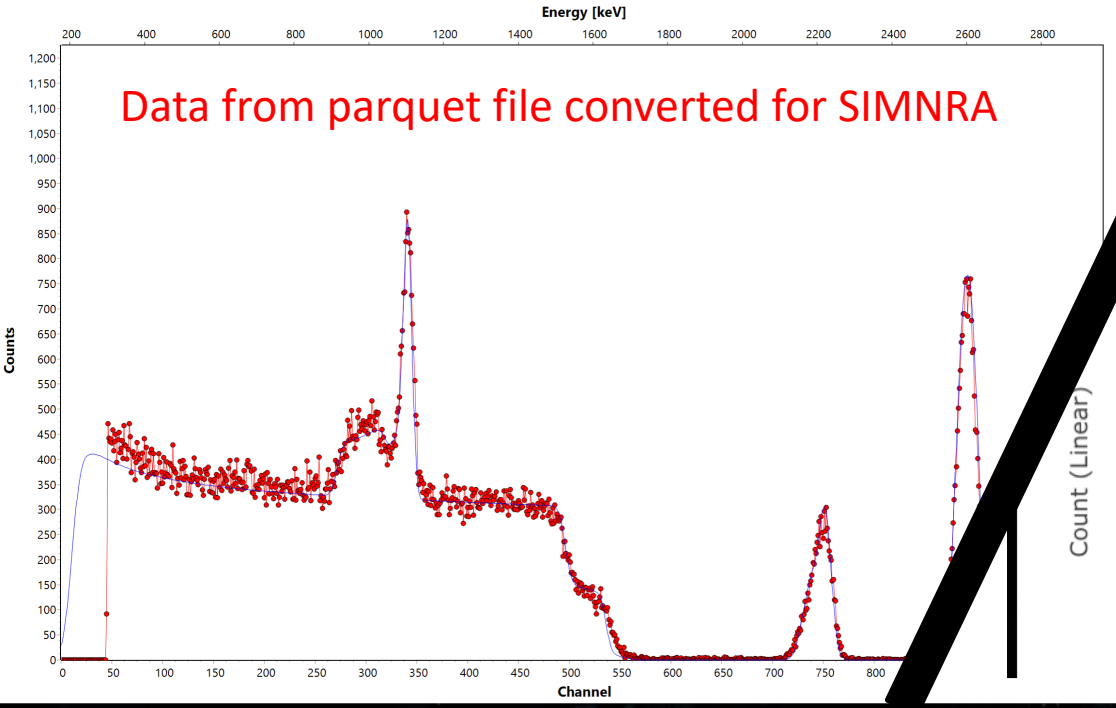
Microservice Implementation (Example)

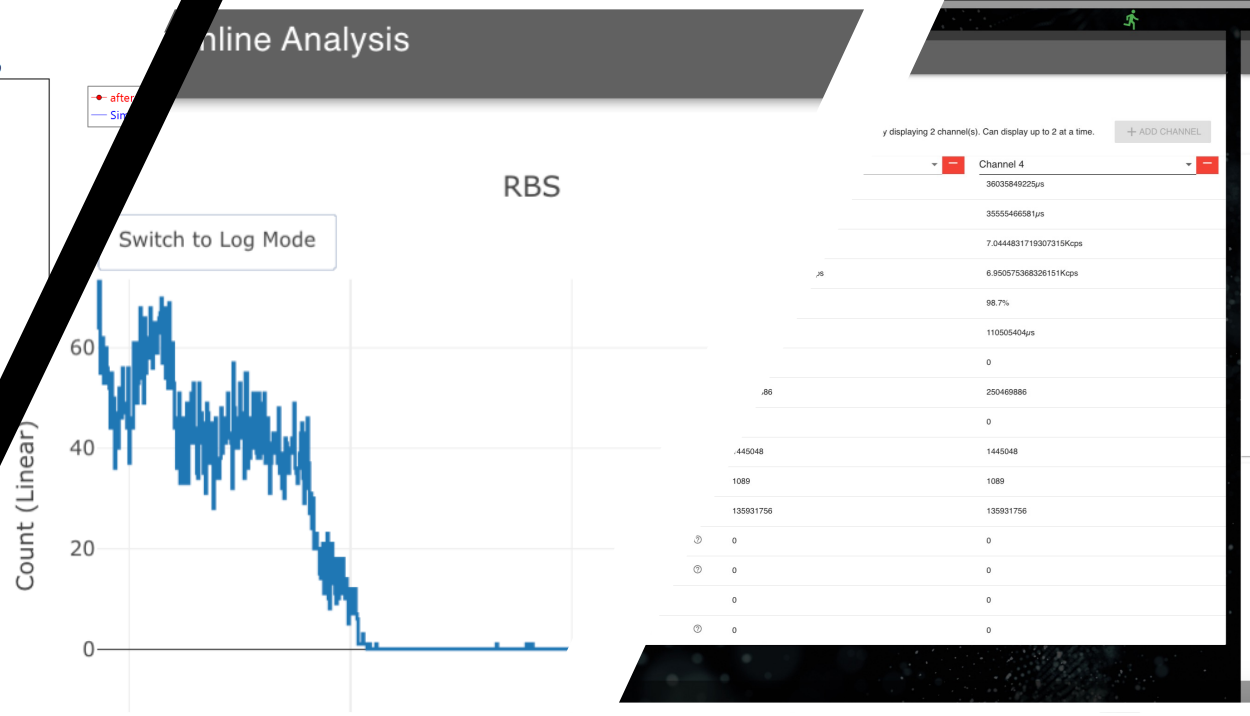


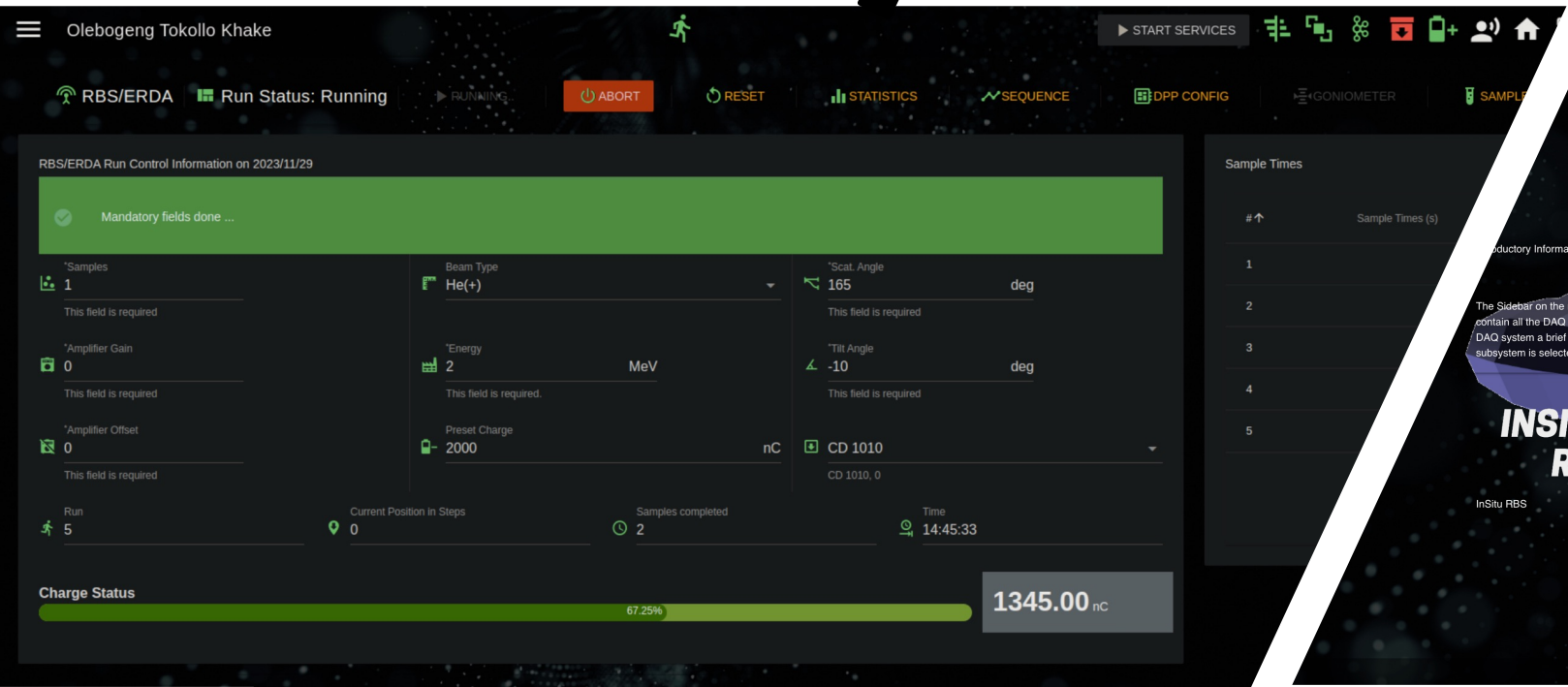
Web UI based on User Requirements

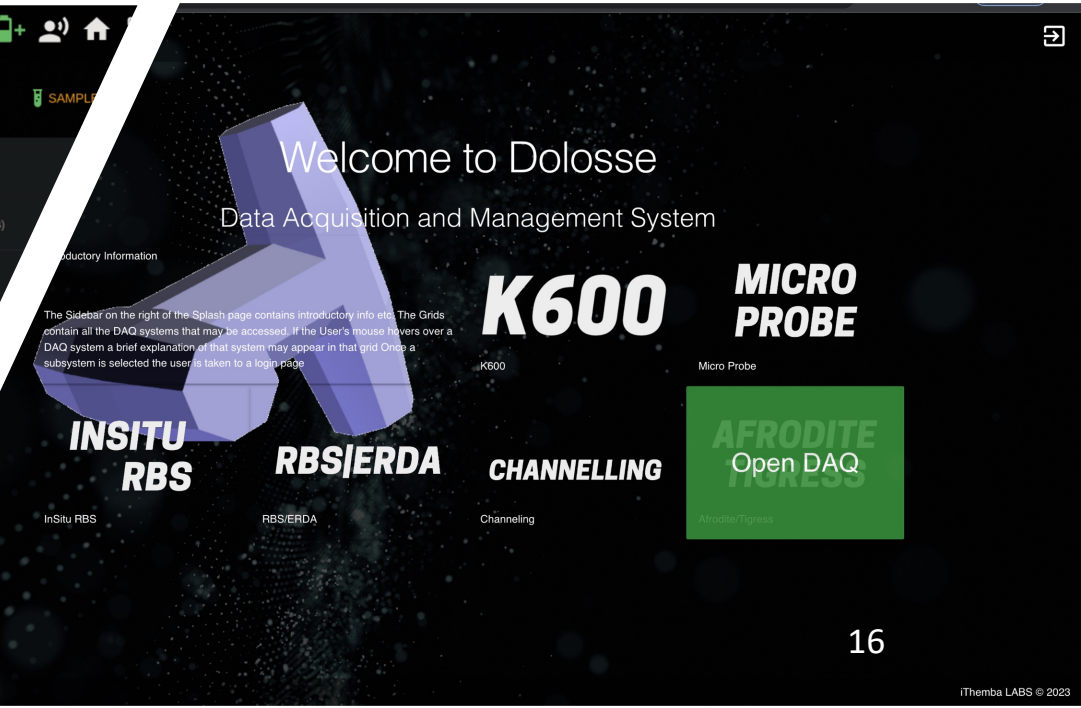


CLI also keeps track of the same information.









iThemba LABS © 2023

Summary

- Released already making an impact to IBA team at iThemba LABS
- To modernize data acquisition and management systems in physics experiments, focusing on efficiency, accuracy, and scalability.
- Overcoming the limitations of outdated technology, which affects system adaptability, data integrity, and support.
- Integration of Kafka for efficient, real-time data management.
- Use of Python with future integration of ROOT for advanced data analysis.
- Adoption of Plotly for interactive and effective data visualization on VUE.js.
- Platform-agnostic with remote access capabilities.
- Implementation of Apache Parquet for efficient data storage.
- Continuous engagement with the open-source community to foster collaboration and innovation.

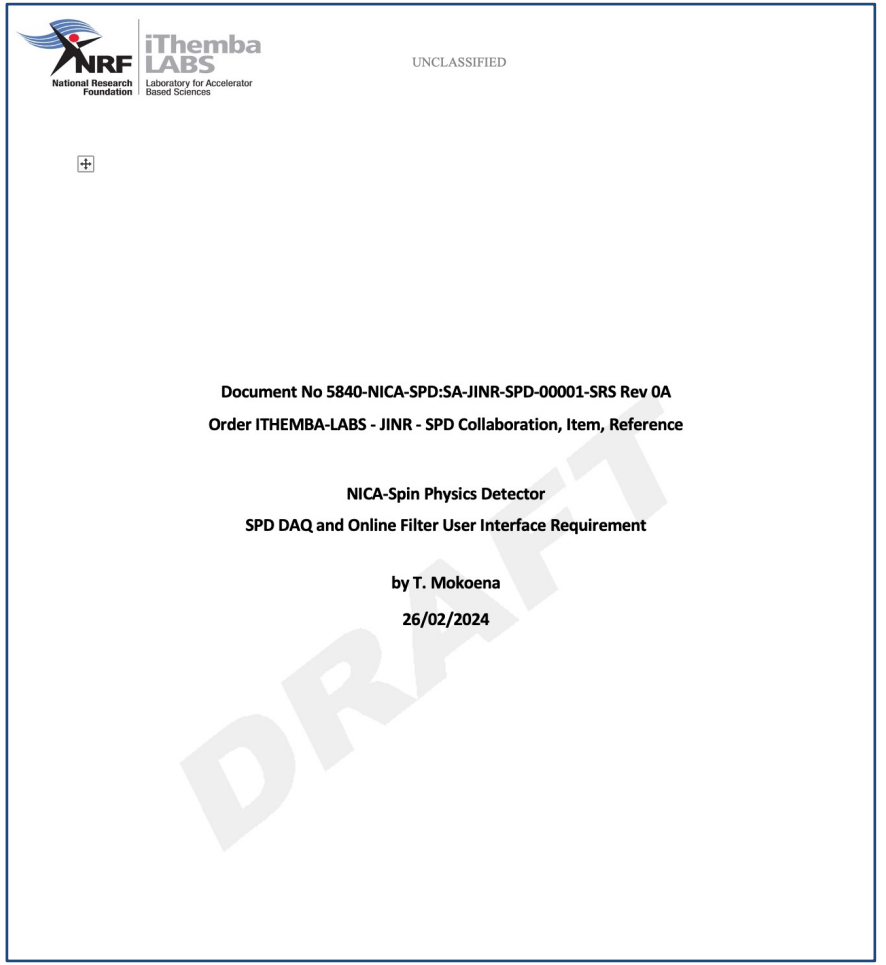
Development Team:
T. Mokoena,
A. Sook,
C. Callaghan,
S. Carelse,
K. Machethe,
O. Khake

Stakeholders –
(Physicists, Nuclear Researchers, Users):
N. Stodart
Dr. Christopher Mtshali,
Dr. Pete Jones,
Dr. Kgashane Malatji,
Dr. Retief Neveling,

Acknowledgement: Dr. S.V. Paulauskas, - Project Science; S. Qhobosheane; C. Peters; B. Losper

SPD Collaboration Plan

- Finalisation of the MoU
- Draft (Revision 0A) of the User Requirement Document shared with the Software and Computing team.
- With main expertise in experimental nuclear physics, readout electronics, software engineering and machine learning techniques, the group plans to participate in the development of the data acquisition and processing for the SPD NICA.
- The primary contributions will be aimed at DAQ design and construction, and Online Filter
- Contribute to the optimization of the data processing, by using standard and machine learning techniques.


NRF National Research Foundation | iThemba LABS Laboratory for Accelerator Based Sciences

UNCLASSIFIED

Document No 5840-NICA-SPD:SA-JINR-SPD-00001-SRS Rev 0A

Order ITHEMBA-LABS - JINR - SPD Collaboration, Item, Reference

NICA-Spin Physics Detector

SPD DAQ and Online Filter User Interface Requirement

by T. Mokoena

26/02/2024

DRAFT

Enkosi, Thank you,
Re a leboga, Siyabonga, Dankie
Ri a livhuwa, Nza khensa

National Research Foundation | SARAO South African Radio Astronomy Observatory

National Research Foundation | SAAO South African Astronomical Observatory

National Research Foundation | SAEON South African Environmental Observation Network

National Research Foundation | SAIAB South African Institute for Aquatic Biodiversity

National Research Foundation | iThemba LABS Laboratory for Accelerator Based Sciences

National Research Foundation | SAASTA South African Agency for Science and Technology Advancement

National Research Foundation | RISA Research and Innovation Support and Advancement