

Dolosse - A Modern, Scalable, and Extensible Data Acquisition and Management System for Nuclear Physics Experiments



24th IEEE Real Time Conference
22 April 2024

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

Dolosse A Modern, Scalable, and Extensible Data Acquisition and Management System for Nuclear Physics Experiments

Agenda

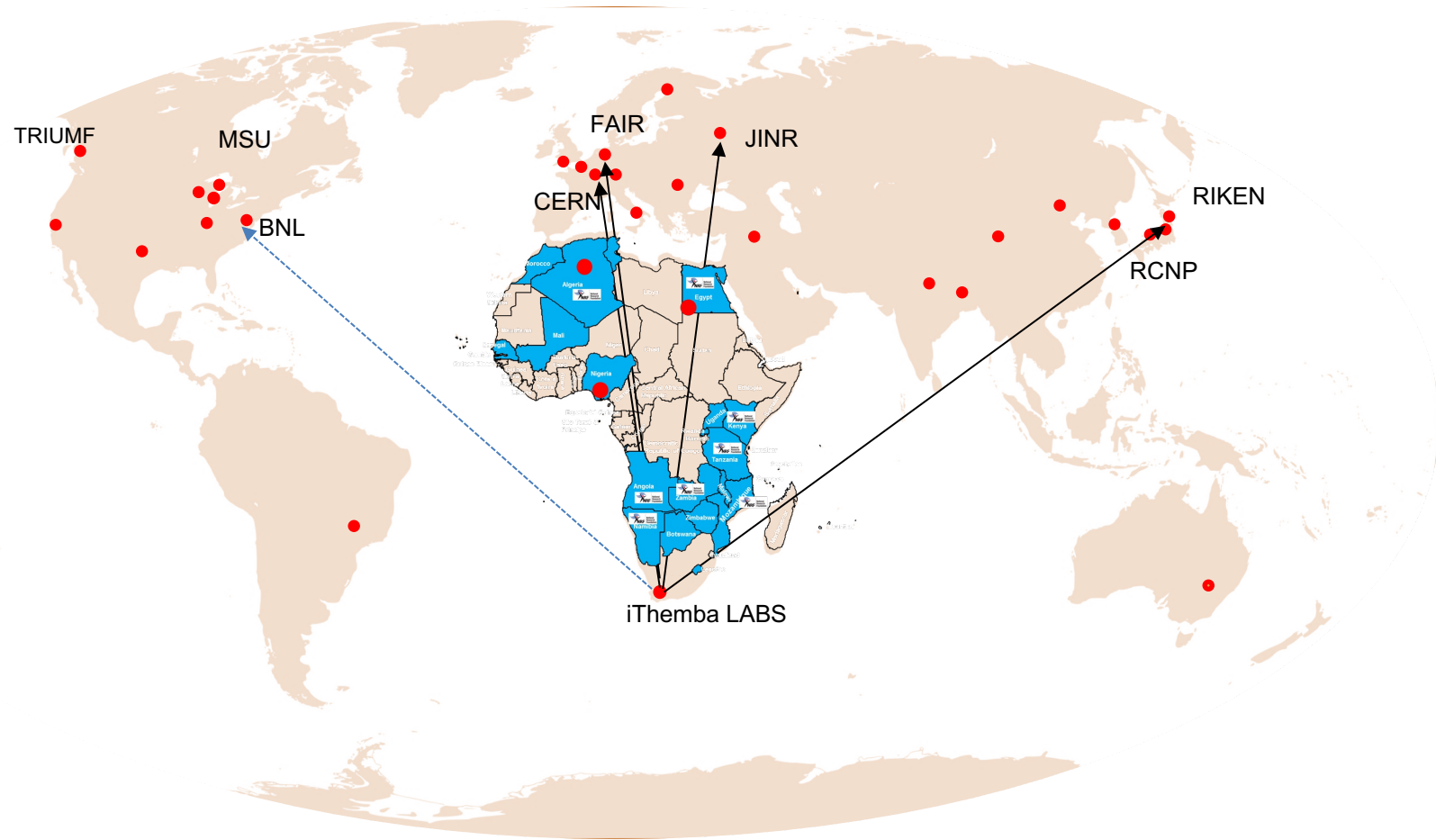
- About iThemba LABS
- Motivation
- Background
- Trends
- Objective
- Dolosse Architecture
- Messaging System
- Storage / Visualization
- Conclusion



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



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

K8 Injector Cyclotron 2



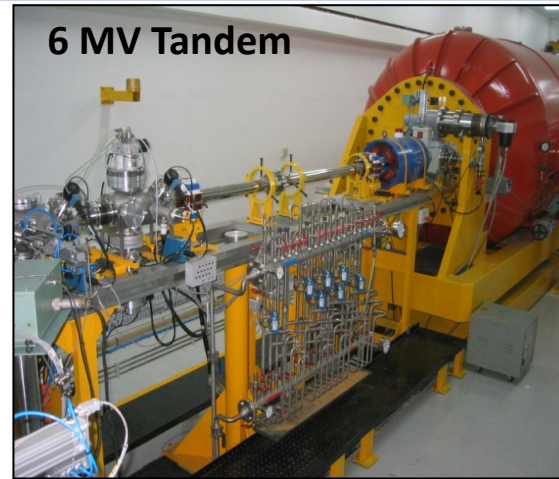
K8 Injector Cyclotron 1



K200 Separated Sector Cyclotron

Subatomic Physics and Applications, Nuclear Medicine

6 MV Tandem



Materials Research and Accelerator Mass Spectrometry (AMS)

K11 Cyclotron



Radioisotope Production



3 MV Tandetron

Materials Research & Nanoscience



IBA C70 Cyclotron

South African Isotope Facility

Motivation: Existing Challenges

Reliance on outdated software and hardware

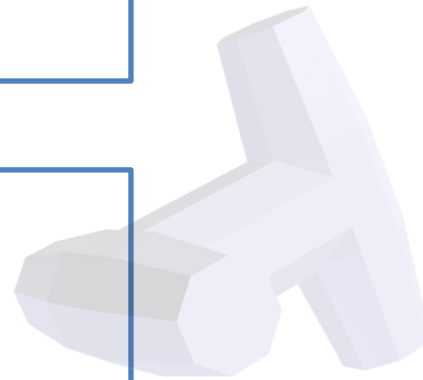
- The **continued use of obsolete technology** restricts advancements and leads to increased maintenance.

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

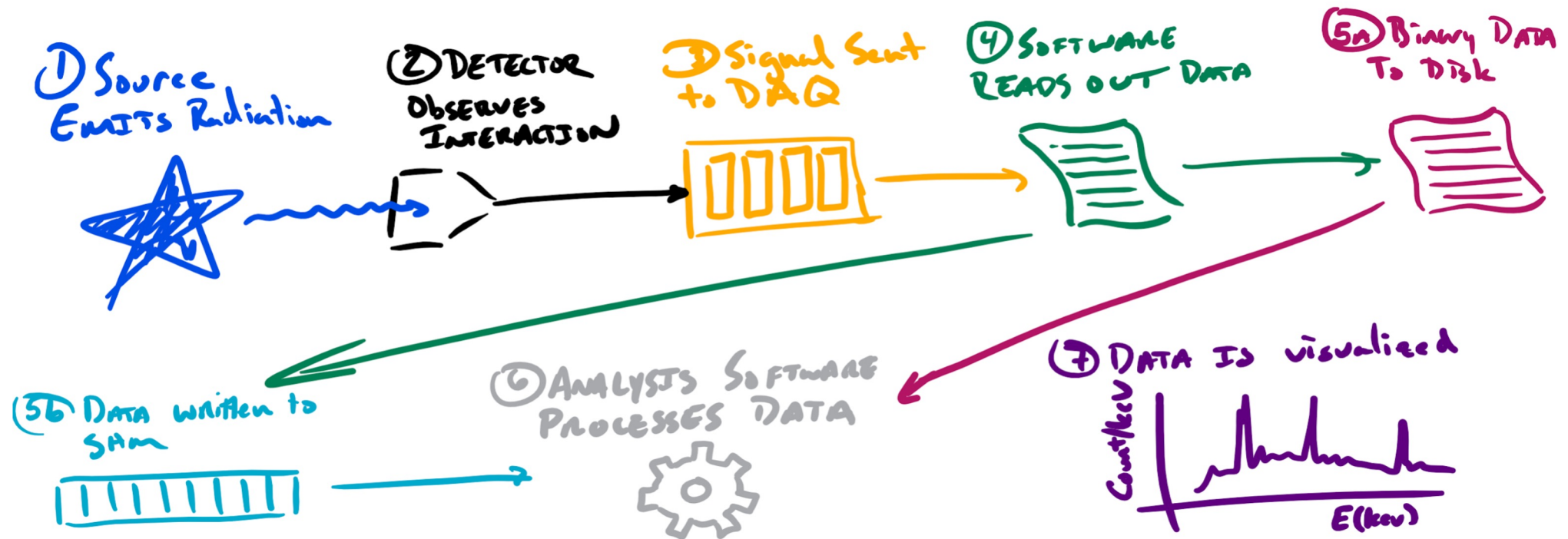


Image source: Dr Stan Paulauskas

- 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

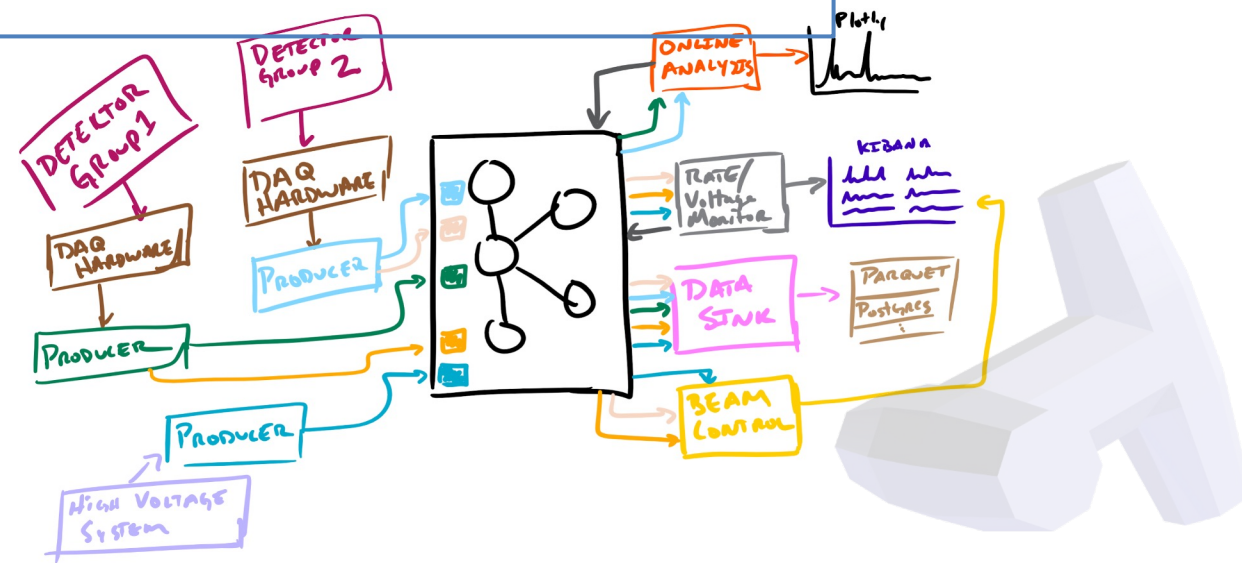
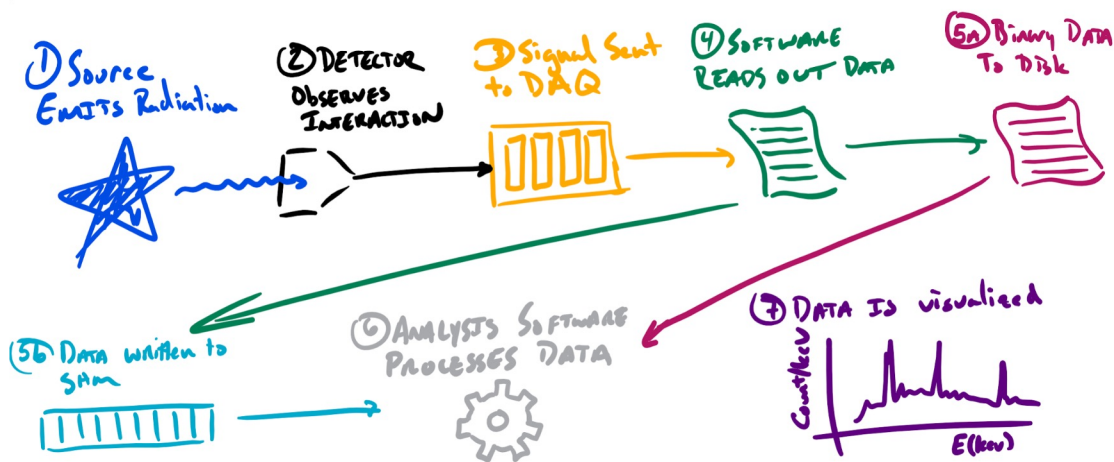
● Industry

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

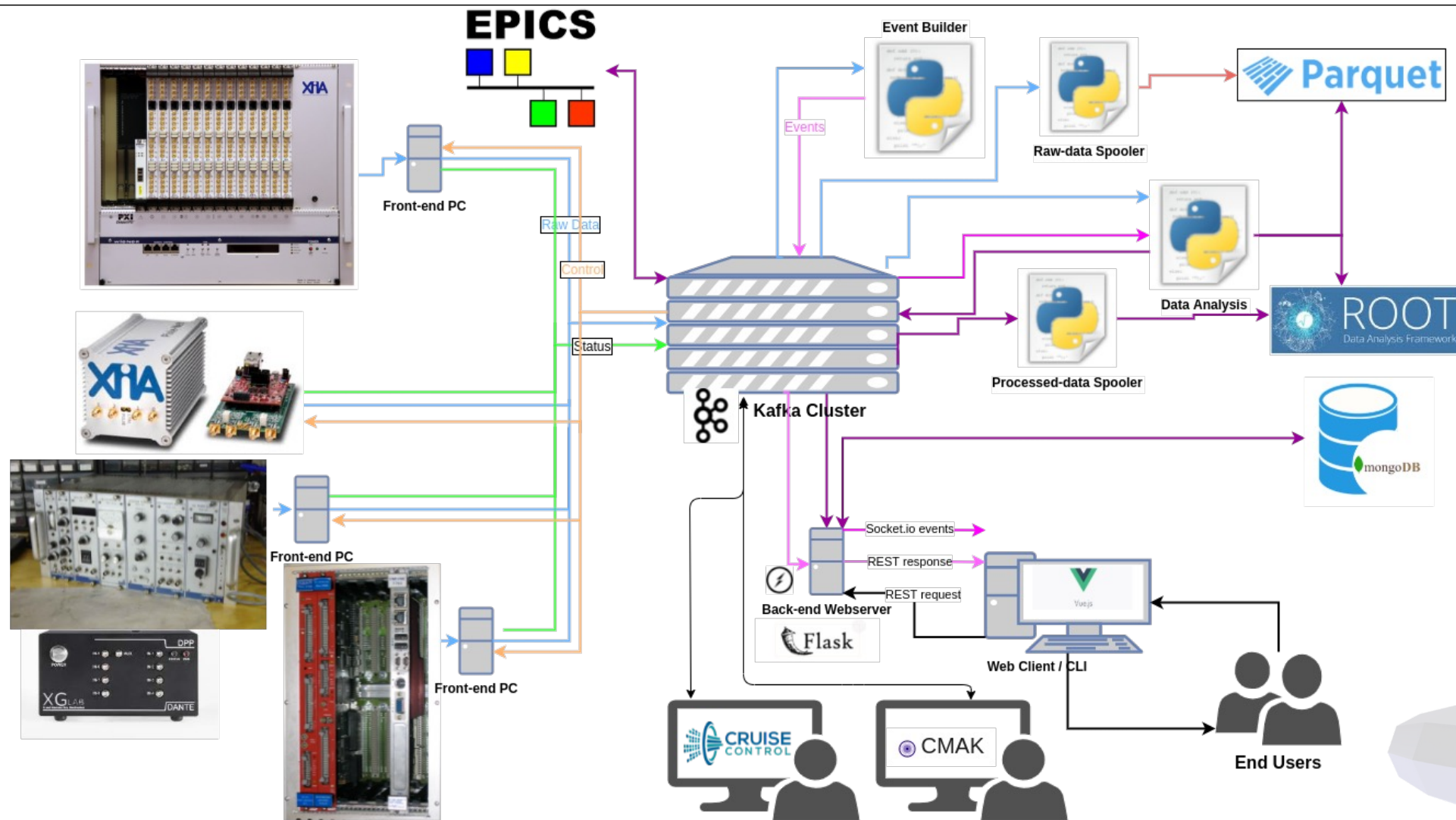


Objective of the Dolosse Project

- 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

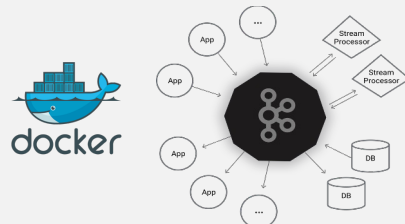


The Approach

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** | 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 | 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

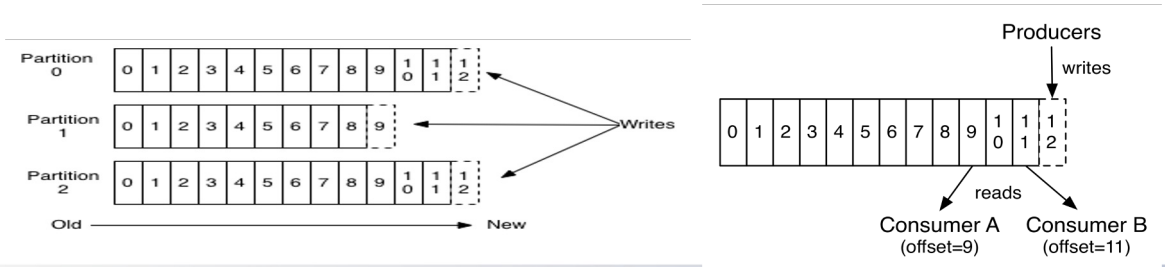
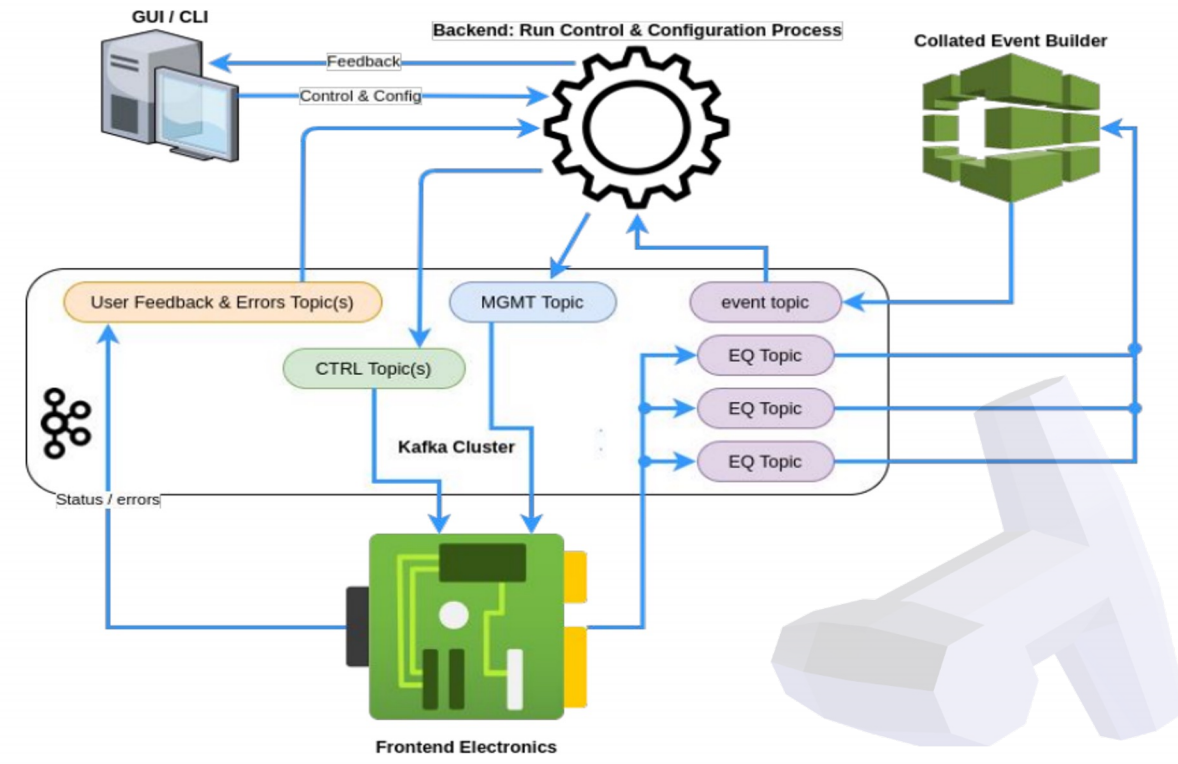
Topics:

- 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.

```
{ "Category": "ID" "Technique": " ID" "Run ID": "int" "Data descriptor": [{"dict}] }
```

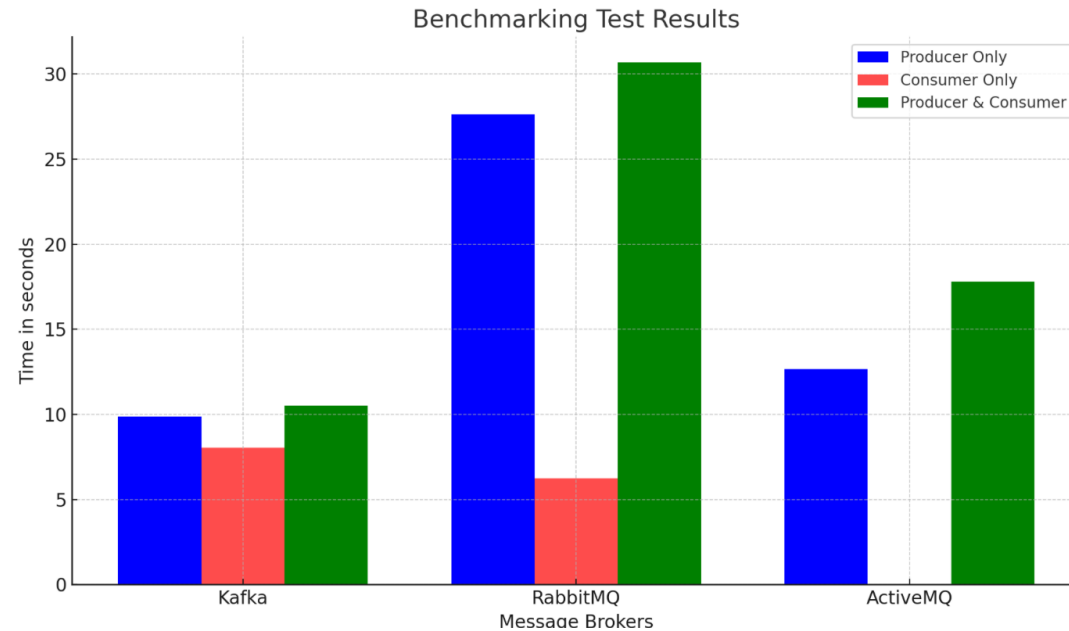
Kafka

Downstream Consumers



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.

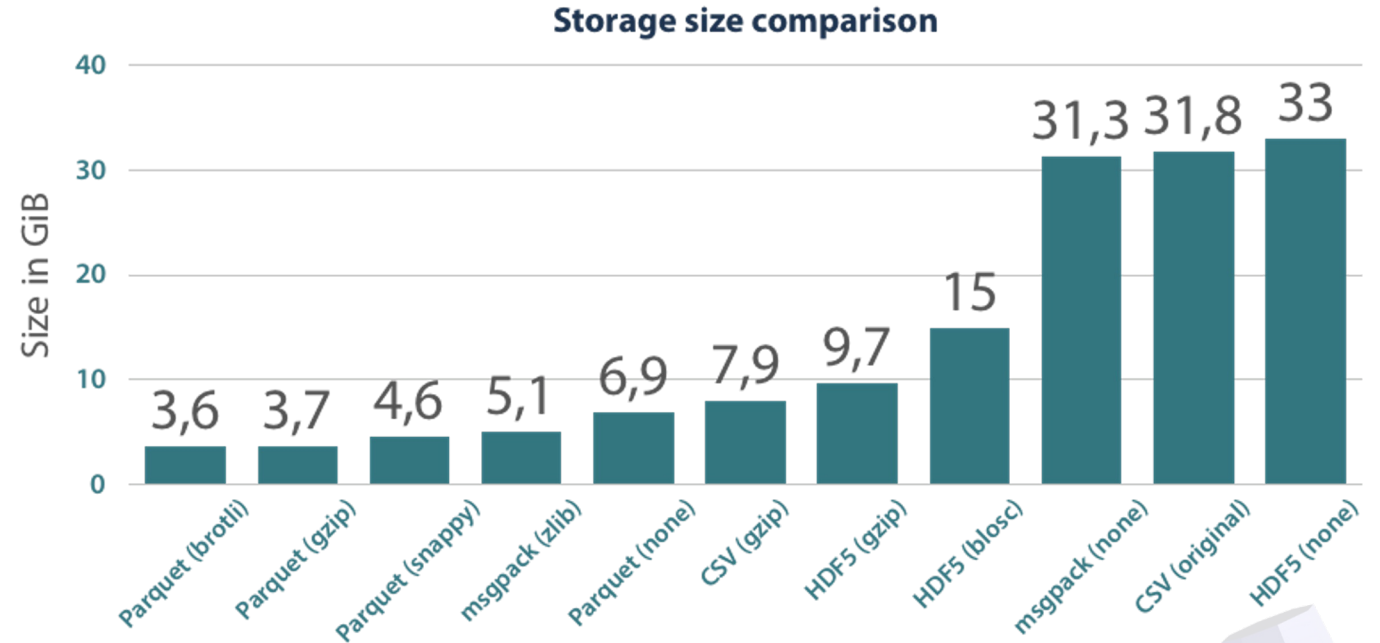


- **RabbitMQ** shows its strength particularly in message consumption speed but **falls behind in production speed and combined operation** efficiency.
- **ActiveMQ** offers a middle ground in production speed but lacks data on consumption speed for a comprehensive comparison. Its combined operation 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.

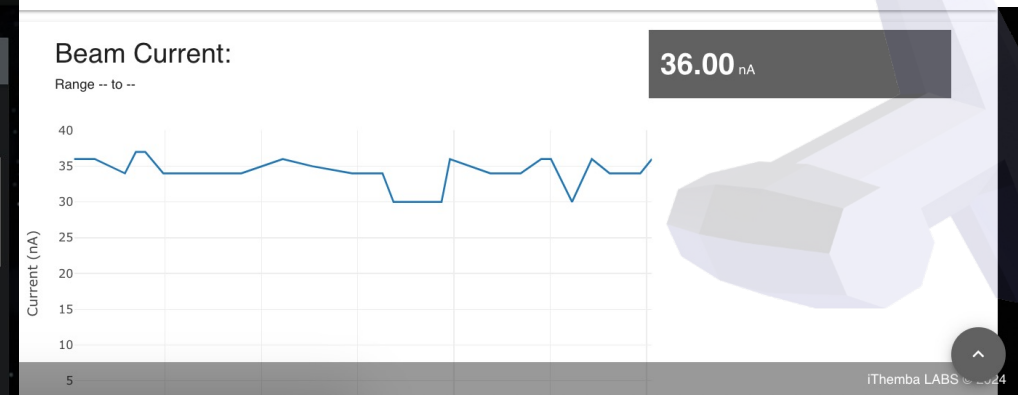
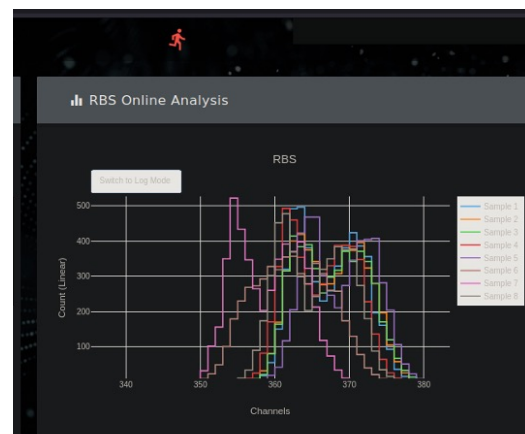
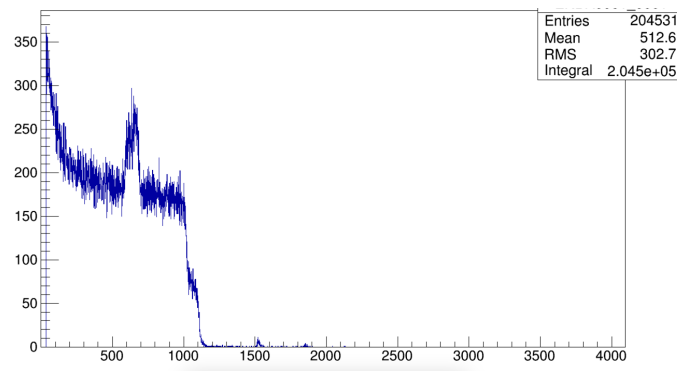
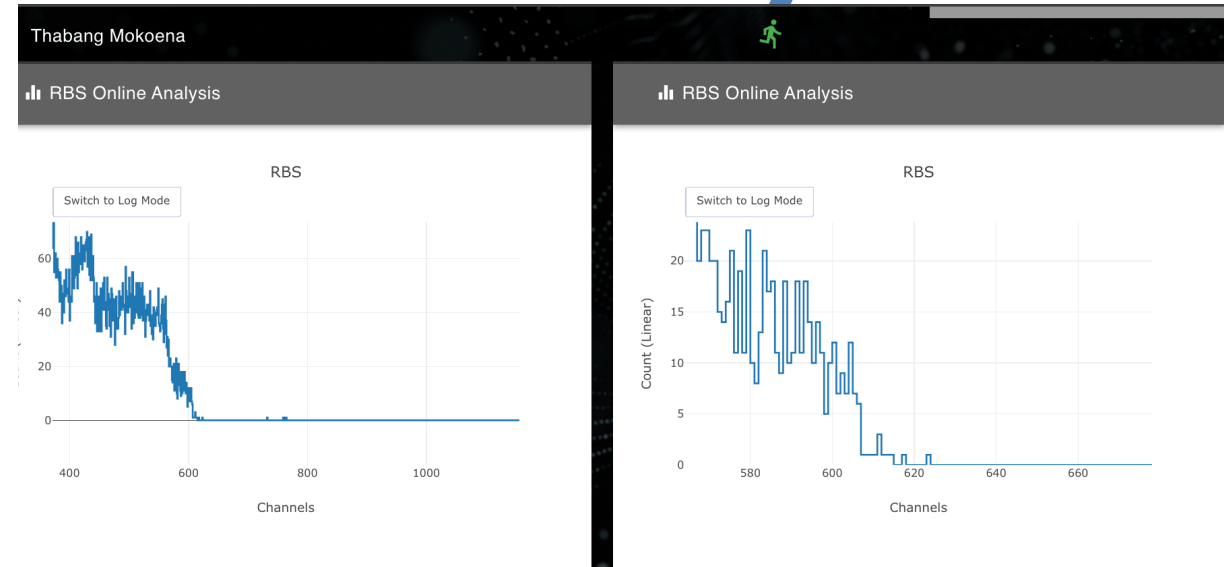


From <https://tech.ida.com/efficient-dataframe-storage-with-apache-parquet/>

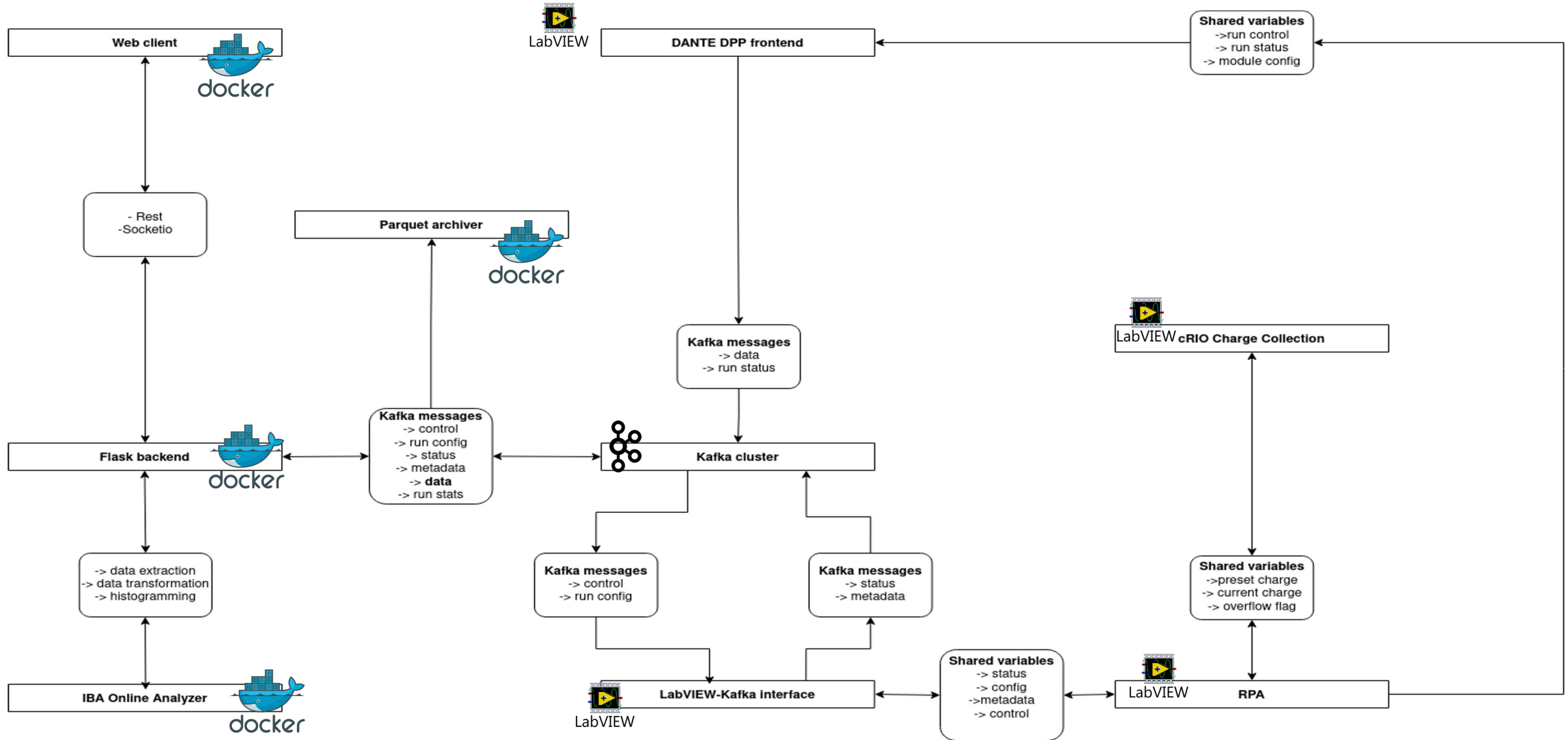
```
master @ caswell ~ gotham @ ~ @ Projects @ dolosse-itl @ dolosse @ feature/342-parquet_write_file @ $ ls
Data Dockerfile.db Dockerfile.rdb dolosse package-lock.json pytest.ini setup.py
docker-compose.yml Dockerfile.epics Dockerfile.wd LICENSE parquet_archiver.log README.md test
Dockerfile.api Dockerfile.evb Documentation package.json parquet_iba_parser.log requirements.txt
master @ caswell ~ gotham @ ~ @ Projects @ dolosse-itl @ dolosse @ feature/342-parquet_write_file @ $
```

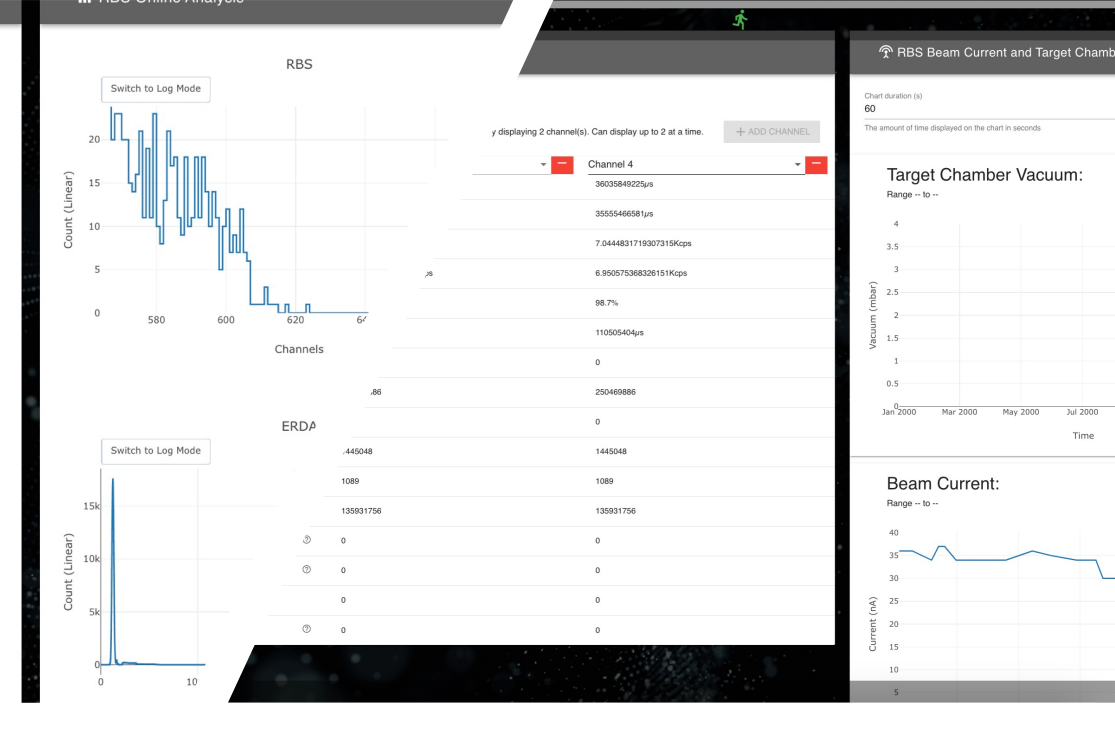
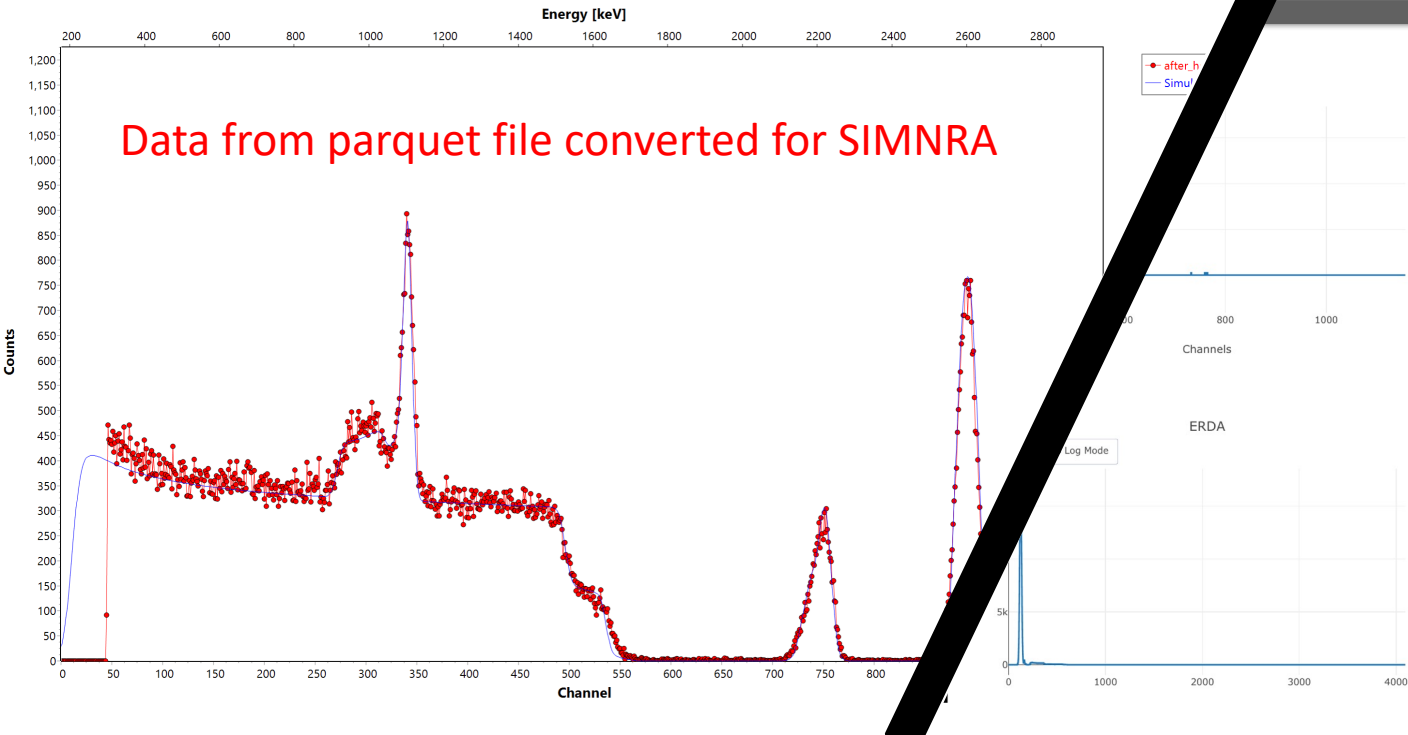

Visualization with Plotly

- 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)





Olebogeng Tokollo Khake

START SERVICES

RBS/ERDA Run Status: Running

ABORT RESET STATISTICS SEQUENCE DPP CONFIG GONIOMETER SAMPLE

RBS/ERDA Run Control Information on 2023/11/29

Mandatory fields done ...

*Samples: 1	Beam Type: He(+)	*Scat. Angle: 165 deg
*Amplifier Gain: 0	*Energy: 2 MeV	*Tilt Angle: -10 deg
*Amplifier Offset: 0	Preset Charge: 2000 nC	CD 1010

Run: 5 | Current Position In Steps: 0 | Samples completed: 2 | Time: 14:45:33

Charge Status: 67.25% | 1345.00 nC

Welcome to Dolosse
Data Acquisition and Management System

INSITU RBS | RBS/ERDA | CHANNELLING | AFRODITE Open DAQ

Sample Times

#	Sample Times (s)
1	
2	
3	
4	
5	

15

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

Acknowledgement

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

Stakeholders –

(Physicists, Nuclear Researchers, Users):

Dr. Christopher Mtshali,
Dr. Pete Jones,
Dr. Kgashane Malatji,
Dr. Retief Neveling,

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



SARAO
South African Radio
Astronomy Observatory



SAAO
South African
Astronomical Observatory



SAEON
South African Environmental
Observation Network



SAIAB
South African Institute
for Aquatic Biodiversity



**iThemba
LABS**
Laboratory for Accelerator
Based Sciences



SAASTA
South African Agency for Science
and Technology Advancement



RISA
Research and Innovation
Support and Advancement