A High-Performance Distributed Architecture for Accelerator Data Processing: The Machine Learning Dataplatform and Dataprovider

Automated Ingestion, Intelligent Querying, and Collaborative Annotation for Accelerator Physics Research

Nicholas Mamais(1,2) – mamais@slac.Stanford.edu, E. Williams(1), C. McChesney(2), C. Allen(2), Kuktae Kim(1), M. Gibbs(1), G. White(1), A. Edelen(1)

1. SLAC National Accelerator Laboratory, USA. 2. Osprey Distributed Control Systems

ABSTRACT

Accelerator facilities generate massive volumes of data that researchers struggle to access and analyze efficiently. We present the implementation of the Machine Learning Data Platform (MLDP) with integrated DataProvider for SLAC accelerator data processing (Figure 1). The system transforms heterogeneous raw accelerator data into ML-ready datasets through systematic ingestion, querying, and annotation workflows, enabling researchers to focus on physics analysis rather than data management challenges.

PHASE 1: DATA INGESTION

Our ingestion system automatically converts different types of accelerator data into a single, organized format. It pulls data from H5 files [1], EPICS archives [2], live streams, and scheduling systems, then enriches it with location and device information before storing it in a searchable database [3] (Figure 2).

PHASE 2: DATA QUERYING

Our query system lets researchers easily find and analyze accelerator data using simple searches, pattern matching, or bulk requests. Users can work with millions of data points using familiar date formats, and the system automatically calculates statistics like averages and data quality metrics without requiring users to download massive datasets (Figure 3).

PHASE 3: ANNOTATION AND ML DATASET PREP.

Our annotation implementation creates organized datasets for machine learning by automatically grouping devices based on their location and function. Researchers can collaborate to tag and categorize data, making it easy to find relevant datasets for analysis. The system exports data in ML-ready formats while tracking where each dataset came from and how devices relate to each other spatially (Figure 4).

CONCLUSIONS

The Machine Learning Data Platform with integrated Data Provider transforms accelerator data management from manual, fragmented workflows into automated, intelligent analysis pipelines. Our three-service architecture demonstrates that systematic spatial enrichment can extract meaningful context from device naming conventions while enabling real-time data discovery across heterogeneous sources. This empowers researchers to correlate experimental data with operational configurations seamlessly, establishing a new paradigm for data-driven accelerator physics research. The platform proves that intelligent data infrastructure can eliminate traditional barriers between raw accelerator data and scientific insight, enabling researchers to focus on physics discovery rather than data wrangling.

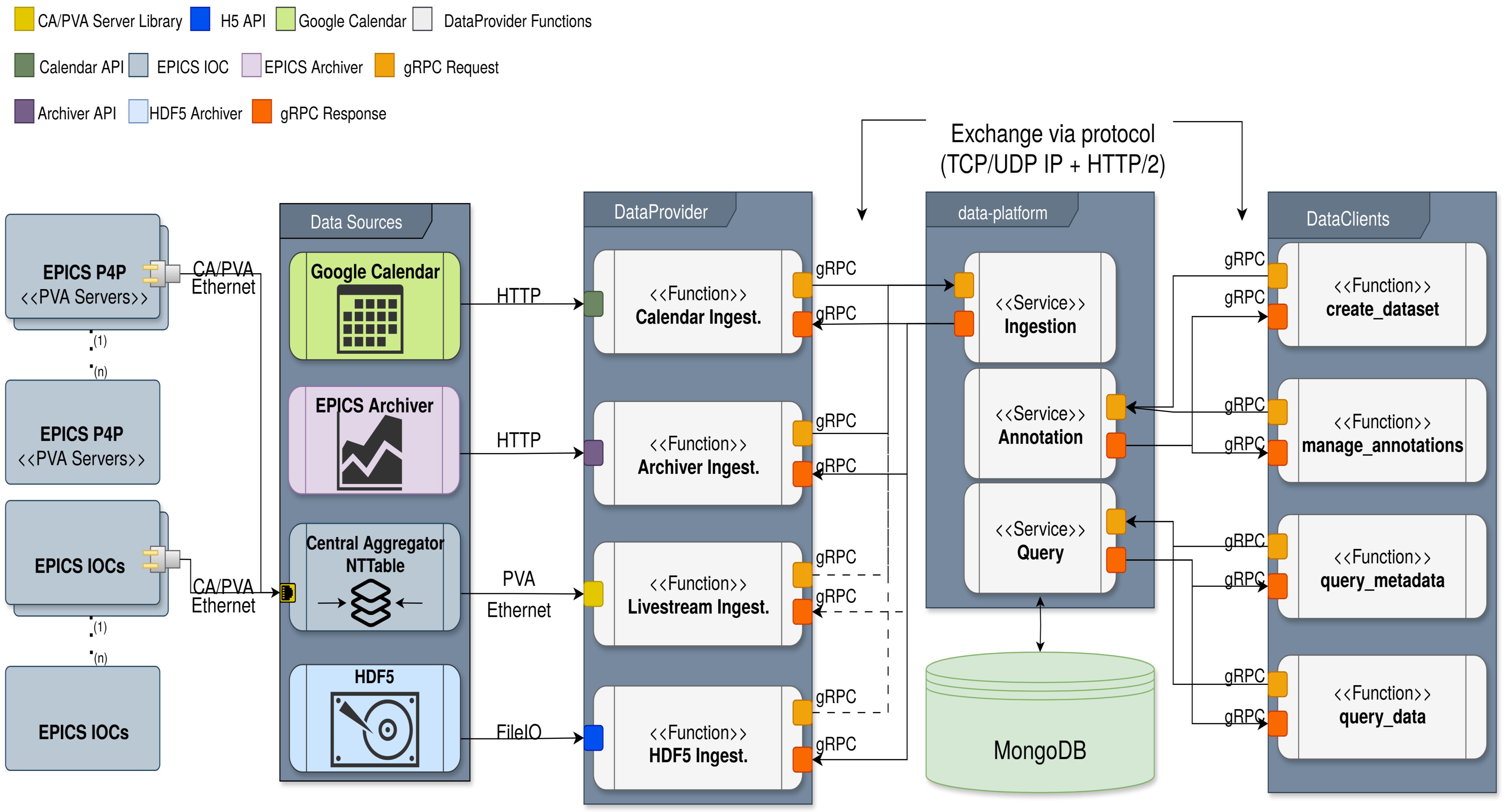
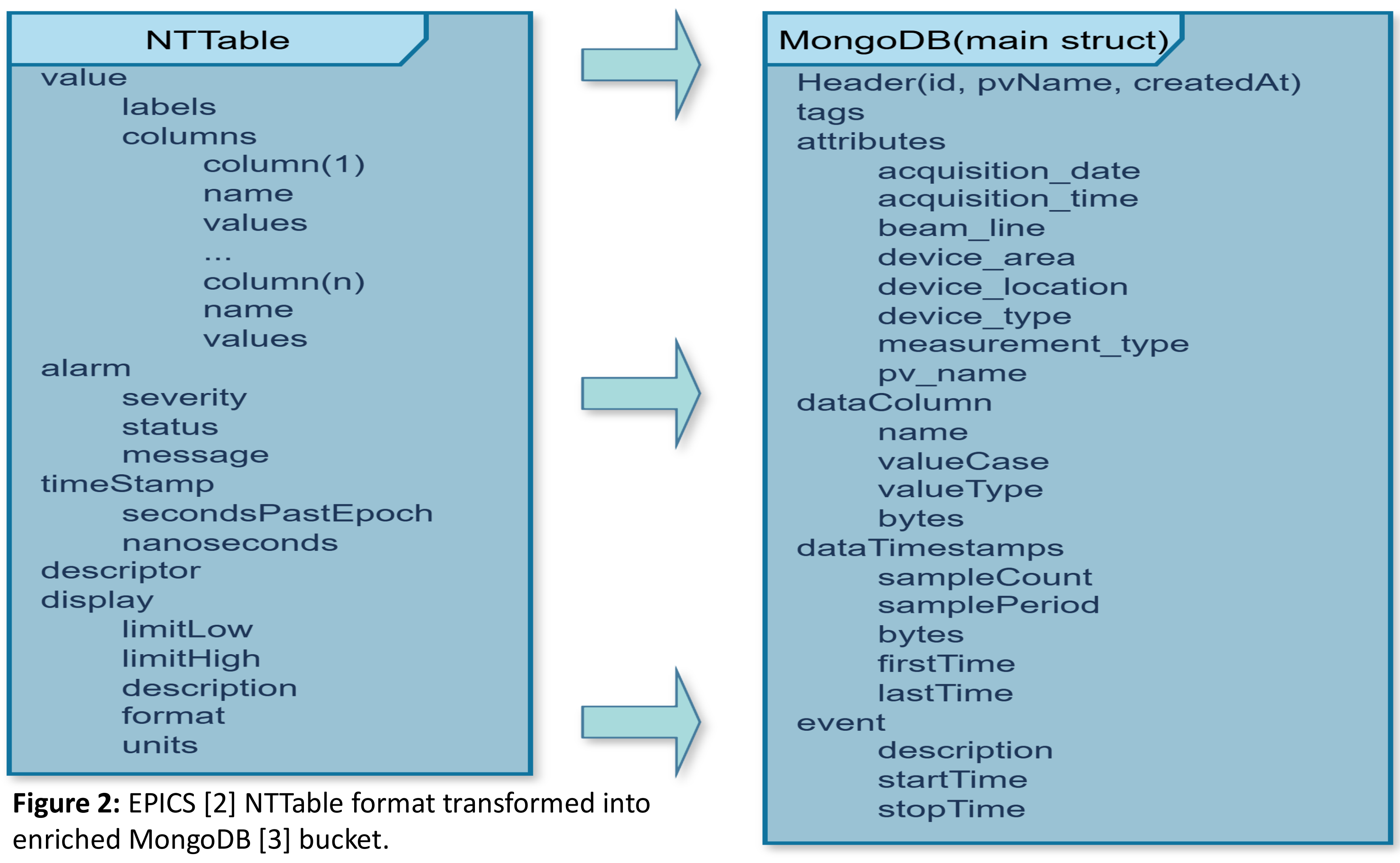


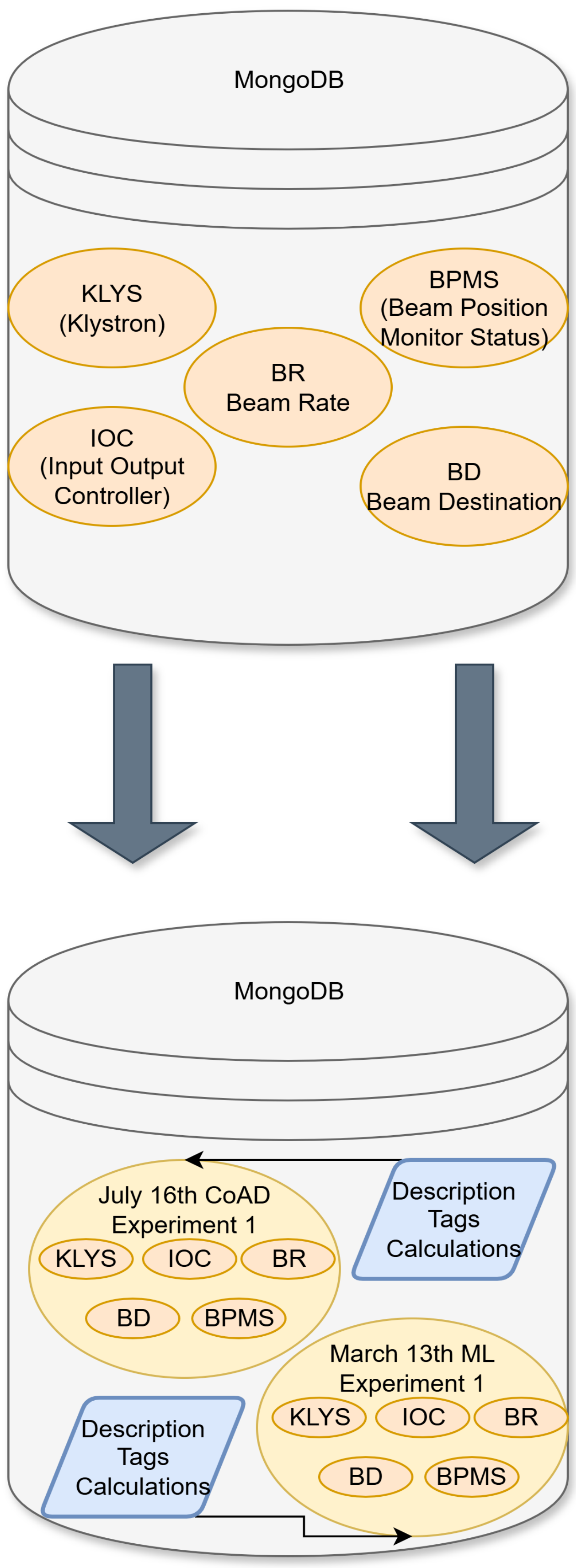
Figure 1: EPICS-based control system architecture showing various data sources flow to the data platform via gRPC protocols and integrated DataProvider

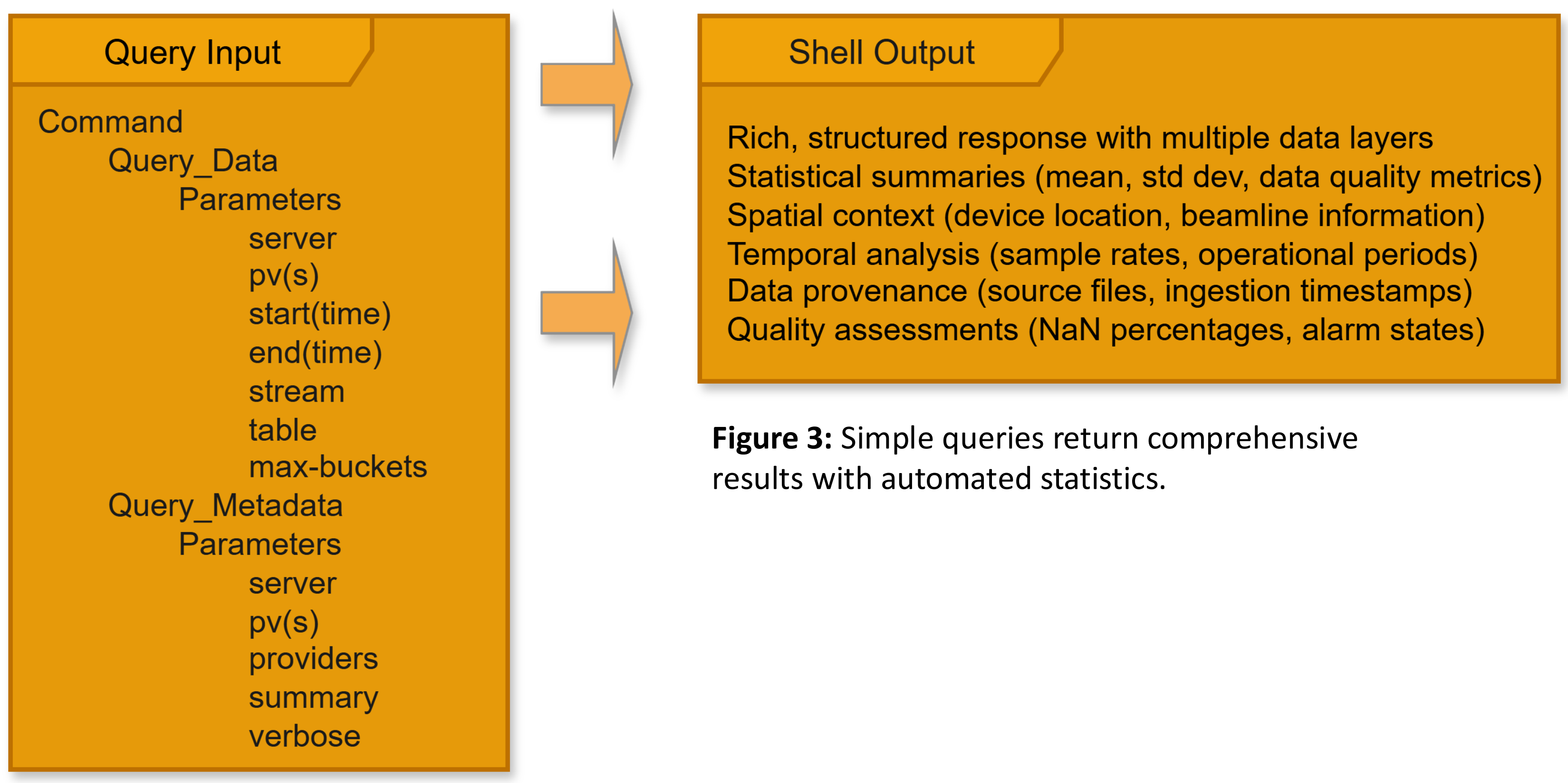


```
NTTable
value
    labels
    columns
        column(1)
        name
        values
        ...
        column(n)
        name
        values
alarm
    severity
    status
    message
timeStamp
    secondsPastEpoch
    nanoseconds
descriptor
display
    limitLow
    limitHigh
    description
    format
    units
```

```
MongoDB(main struct)
Header(id, pvName, createdAt)
tags
attributes
    acquisition_date
    acquisition_time
    beam_line
    device_area
    device_location
    device_type
    measurement_type
    pv_name
dataColumn
    name
    valueCase
    valueType
    bytes
dataTimestamps
    sampleCount
    samplePeriod
    bytes
    firstTime
    lastTime
event
    description
    startTime
    stopTime
```

Figure 2: EPICS [2] NTTable format transformed into enriched MongoDB [3] bucket.





```
Query Input
Command
    Query_Data
        Parameters
            server
            pv(s)
            start(time)
            end(time)
            stream
            table
            max-buckets
    Query_Metadata
        Parameters
            server
            pv(s)
            providers
            summary
            verbose
```

Shell Output

Rich, structured response with multiple data layers
Statistical summaries (mean, std dev, data quality metrics)
Spatial context (device location, beamline information)
Temporal analysis (sample rates, operational periods)
Data provenance (source files, ingestion timestamps)
Quality assessments (NaN percentages, alarm states)

Figure 3: Simple queries return comprehensive results with automated statistics.

Figure 4: Scattered datasets organized into tagged, ML-ready collections.

ACKNOWLEDGMENTS

The authors thank Osprey Distributed Control Systems for developing the MLDP architecture and SLAC National Accelerator Laboratory for providing accelerator data and testing environment. Special recognition to the LCLS-II operations team for technical requirements and testing support.

REFERENCES

[1] The HDF Group (2024). HDF5 Scientific Data Format. https://www.hdfgroup.org
[2] EPICS Collaboration (2024). Experimental Physics and Industrial Control System. https://epics.anl.gov
[3] MongoDB Inc. (2024). MongoDB Document Database. https://www.mongodb.com