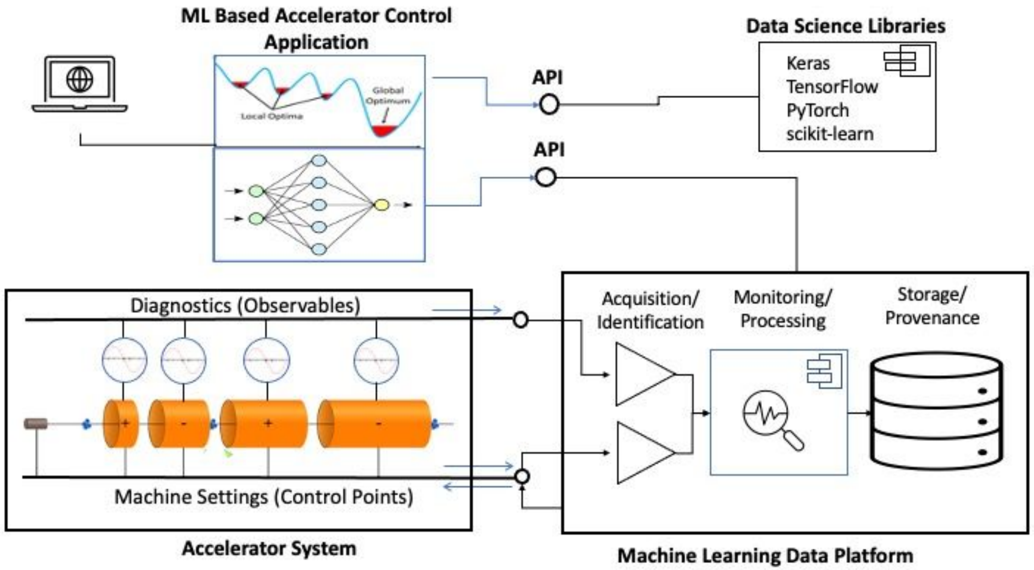
Machine Learning Data Platform By Example

Craig McChesney
Christopher K. Allen
Mitch Frauenheim
Osprey Scientific Control Systems

Osprey
Distributed Control Systems

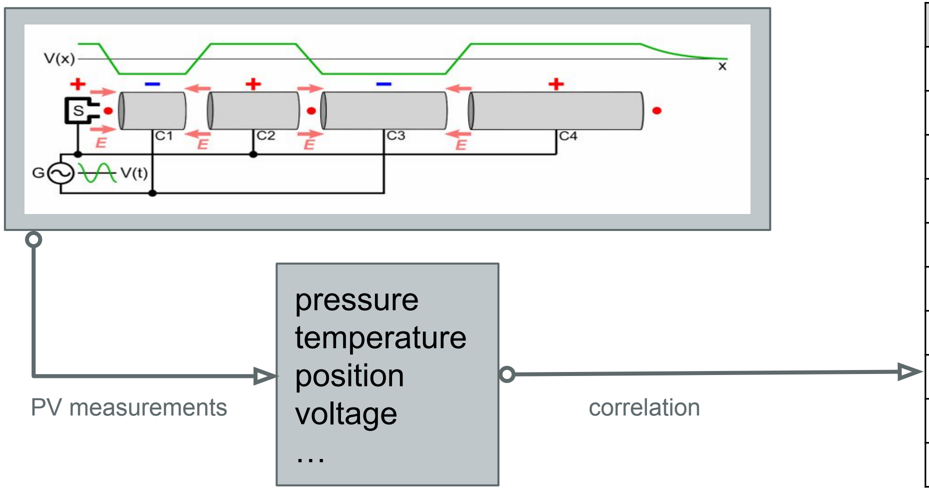
MLDP Motivation

The Machine Learning Data Platform (MLDP) provides full-stack support for machine learning and data science applications for the diagnosis, modeling, control, and optimization of particle accelerator facilities.


ML Based Accelerator Control Application
Local Optima
Global Optimum
API
API
Data Science Libraries
Keras
TensorFlow
PyTorch
scikit-learn
Diagnostics (Observables)
Machine Settings (Control Points)
Accelerator System
Acquisition/ Identification
Monitoring/ Processing
Storage/ Provenance
Machine Learning Data Platform

MLDP Context

The MLDP is intended to be used in a particle accelerator or experimental physics research facility with a control system. Process variables are sampled and correlated in time.



pressure
temperature
position
voltage
…

PV measurements

correlation

timestamp	S01-GCC-1	S01A-BPM	CAMERA-1
12:00:00.000	pressure1	[x1, y1]	image1
12:00:00.250	pressure2	[x2, y2]	image2
12:00:00.500	pressure3	[x3, y3]	image3
12:00:00.750	pressure4	[x4, y4]	image4
12:00:01.000	pressure5	[x5, y5]	image5
12:00:01.250	pressure6	[x6, y6]	image6
12:00:01.500	pressure7	[x7, y7]	image7
12:00:01.750	pressure8	[x8, y8]	image8
12:00:02.000	pressure9	[x9, y9]	image9
12:00:02.250	pressure10	[x10, y10]	image10

(image with permission from Chetvorno, CC0, via Wikimedia)Commons

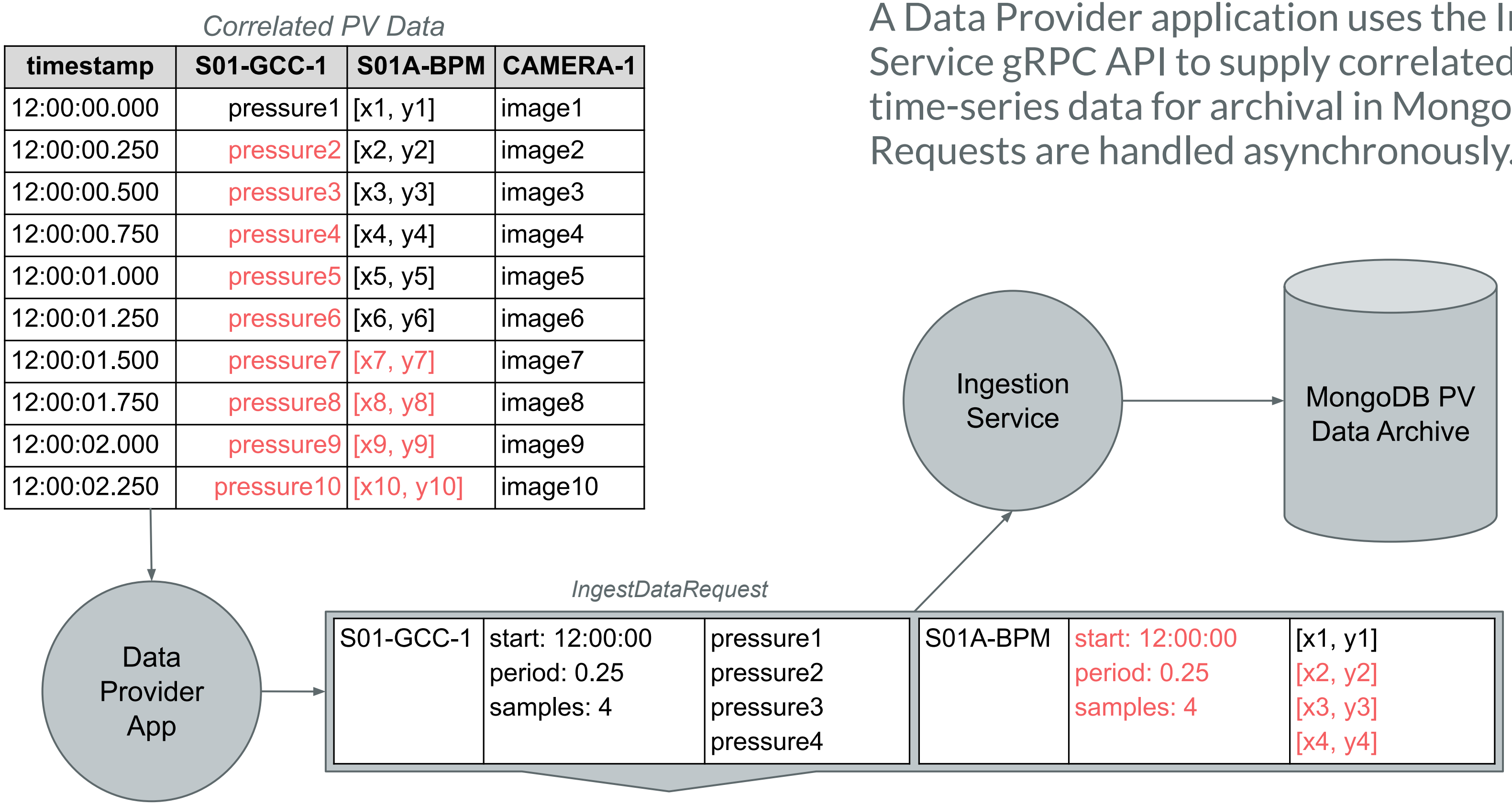
MLDP By Example

Because it is comprised of server applications, an API, and client libraries, it is difficult to give a live MLDP demonstration. The purpose of this presentation is to illustrate MLDP capabilities by way of some simple examples, including:

- time-series data ingestion
- time-series data query
- data provider registration
- ingestion stream error detection
- data provider metadata and PV metadata query
- archive annotation dataset administration
- archive annotation
- calculations upload and provenance tracking
- archive annotation query
- data and calculations export
- ingestion stream subscription

Time-Series Data Ingestion

A Data Provider application uses the Ingestion Service gRPC API to supply correlated time-series data for archival in MongoDB. Requests are handled asynchronously.



Correlated PV Data

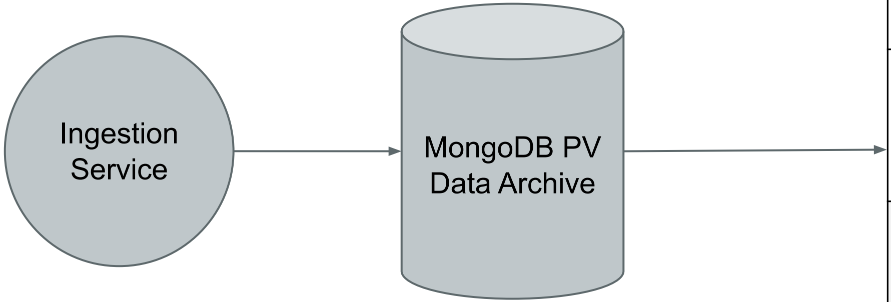
timestamp	S01-GCC-1	S01A-BPM	CAMERA-1
12:00:00.000	pressure1	[x1, y1]	image1
12:00:00.250	pressure2	[x2, y2]	image2
12:00:00.500	pressure3	[x3, y3]	image3
12:00:00.750	pressure4	[x4, y4]	image4
12:00:01.000	pressure5	[x5, y5]	image5
12:00:01.250	pressure6	[x6, y6]	image6
12:00:01.500	pressure7	[x7, y7]	image7
12:00:01.750	pressure8	[x8, y8]	image8
12:00:02.000	pressure9	[x9, y9]	image9
12:00:02.250	pressure10	[x10, y10]	image10

IngestDataRequest

S01-GCC-1	start: 12:00:00 period: 0.25 samples: 4	pressure1 pressure2 pressure3 pressure4	S01A-BPM	start: 12:00:00 period: 0.25 samples: 4	[x1, y1] [x2, y2] [x3, y3] [x4, y4]

Time-Series Data Archive

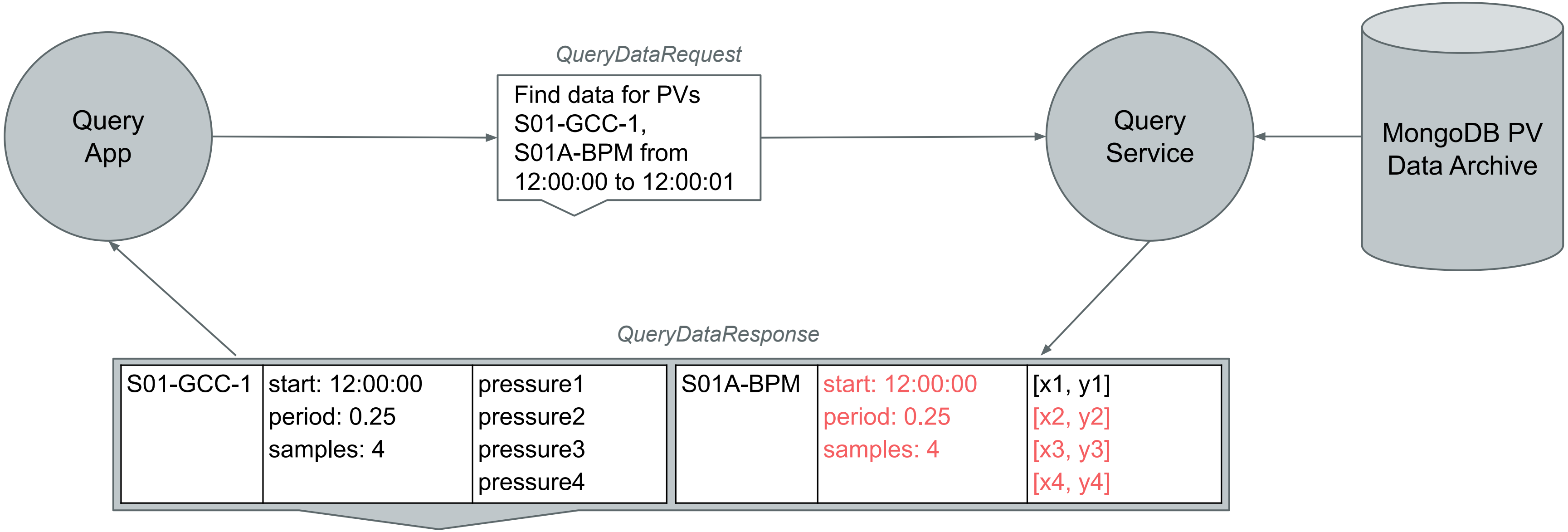
The Ingestion Service uses MongoDB to manage the time-series data archive. For database efficiency, data are organized in "buckets", each containing a vector of heterogeneous sample values for the time period specified by the bucket's "sampling clock" with start time, sample period, and count.



PV	sampling clock	data values
S01-GCC-1	start: 12:00:00 period: 0.25 samples: 4	pressure1 pressure2 pressure3 pressure4
S01-GCC-1	start: 12:00:01 period: 0.25 samples: 4	pressure5 pressure6 pressure7 pressure8
S01A-BPM	start: 12:00:00 period: 0.25 samples: 4	[x1, y1] [x2, y2] [x3, y3] [x4, y4]
S01A-BPM	start: 12:00:01 period: 0.25 samples: 4	[x5, y5] [x6, y6] [x7, y7] [x8, y8]
CAMERA-1	start: 12:00:00 period: 0.25 samples: 4	image1 image2 image3 image4
CAMERA-1	start: 12:00:01 period: 0.25 samples: 4	image5 image6 image7 image8

Time-Series Data Query

A query application sends API request with list of PV names and time range to Query Service, which responds with "data buckets" matching the search criteria.

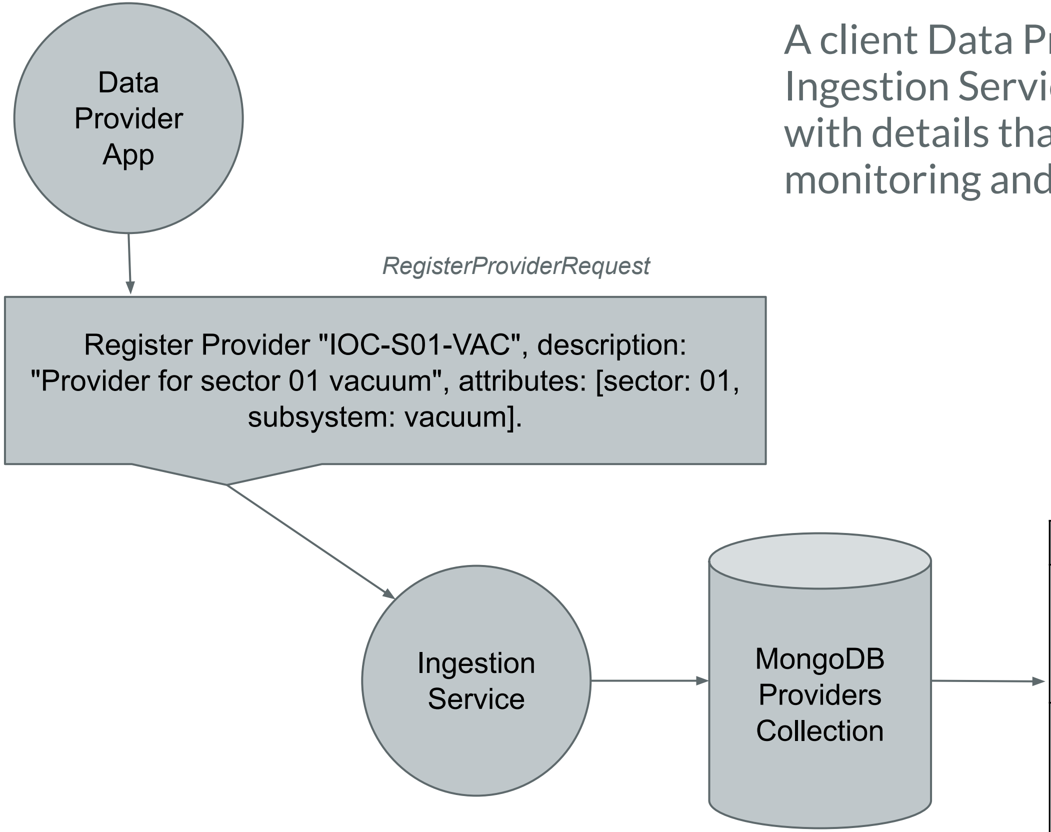


Find data for PVs S01-GCC-1, S01A-BPM from 12:00:00 to 12:00:01

S01-GCC-1	start: 12:00:00 period: 0.25 samples: 4	pressure1 pressure2 pressure3 pressure4	S01A-BPM	start: 12:00:00 period: 0.25 samples: 4	[x1, y1] [x2, y2] [x3, y3] [x4, y4]

Data Provider Registration

A client Data Provider application uses the Ingestion Service Provider API to register itself with details that can be used for ingestion monitoring and provider metadata queries.



Data Provider App

RegisterProviderRequest

Register Provider "IOC-S01-VAC", description: "Provider for sector 01 vacuum", attributes: [sector: 01, subsystem: vacuum].

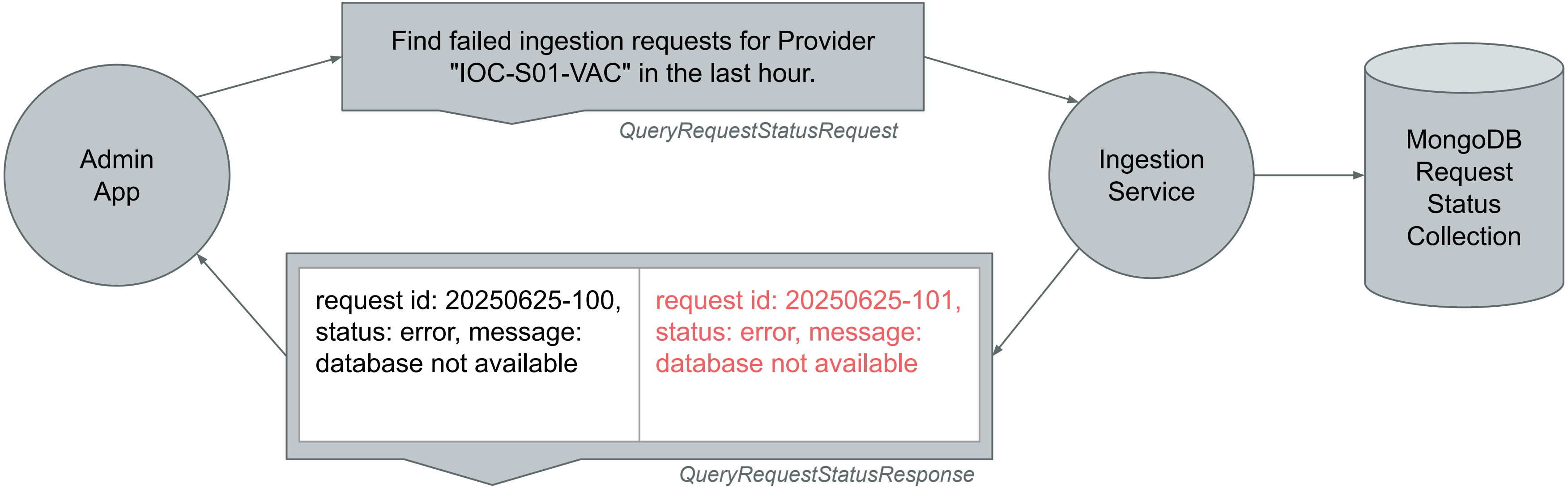
Ingestion Service

MongoDB Providers Collection

Provider	description	attributes
IOC-S01-VAC	Provider for sector 01 vacuum	[sector: 01, subsystem: vacuum]
IOC-S02-VAC	Provider for sector 02 vacuum	[sector: 02, subsystem: vacuum]

Ingestion Stream Error Detection

Data ingestion is performed asynchronously in order to maximize performance. The status of individual ingestion requests is recorded in a database which administrative tools can use to detect problems in data ingestion.



Find failed ingestion requests for Provider "IOC-S01-VAC" in the last hour.

QueryRequestStatusRequest

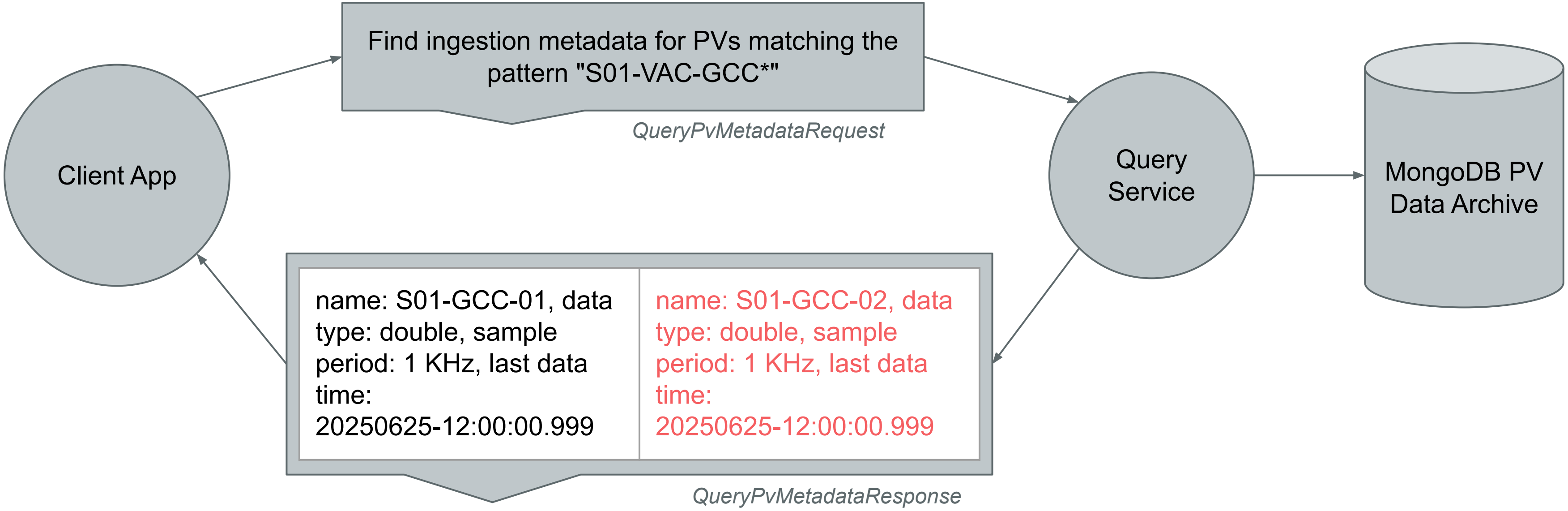
request id: 20250625-100, status: error, message: database not available

request id: 20250625-101, status: error, message: database not available

QueryRequestStatusResponse

PV Metadata Query

A client application uses the Query Service's PV Metadata Query API to find ingestion metadata for PV names matching a specified pattern.



Client App

Find ingestion metadata for PVs matching the pattern "S01-VAC-GCC*"

QueryPvMetadataRequest

Query Service

MongoDB PV Data Archive

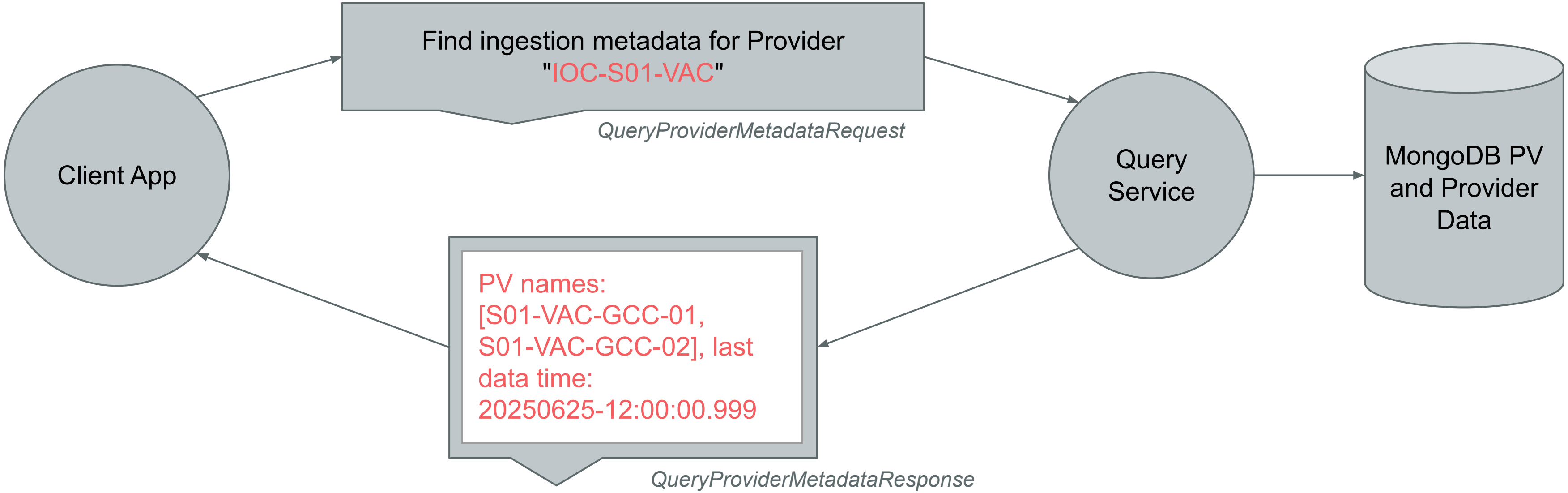
name: S01-GCC-01, data type: double, sample period: 1 KHz, last data time: 20250625-12:00:00.999

name: S01-GCC-02, data type: double, sample period: 1 KHz, last data time: 20250625-12:00:00.999

QueryPvMetadataResponse

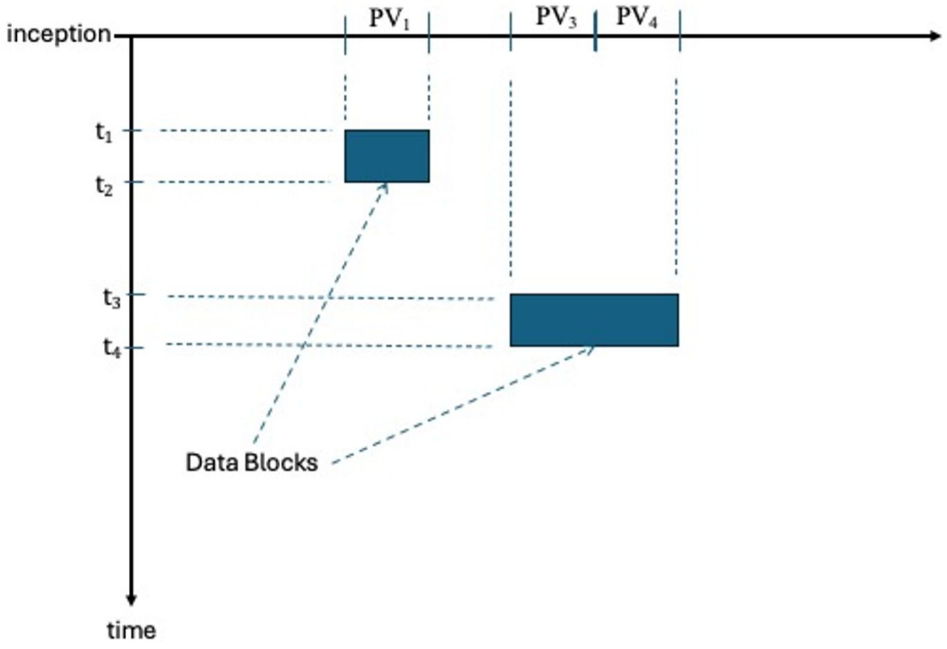
Data Provider Metadata Query

A client application uses the Query Service's Data Provider Metadata Query API to find metadata for data Providers (suppliers of ingestion data).


Find ingestion metadata for Provider "IOC-S01-VAC"
QueryProviderMetadataRequest
Client App
Query Service
MongoDB PV and Provider Data
PV names: [S01-VAC-GCC-01, S01-VAC-GCC-02], last data time: 20250625-12:00:00.999
QueryProviderMetadataResponse

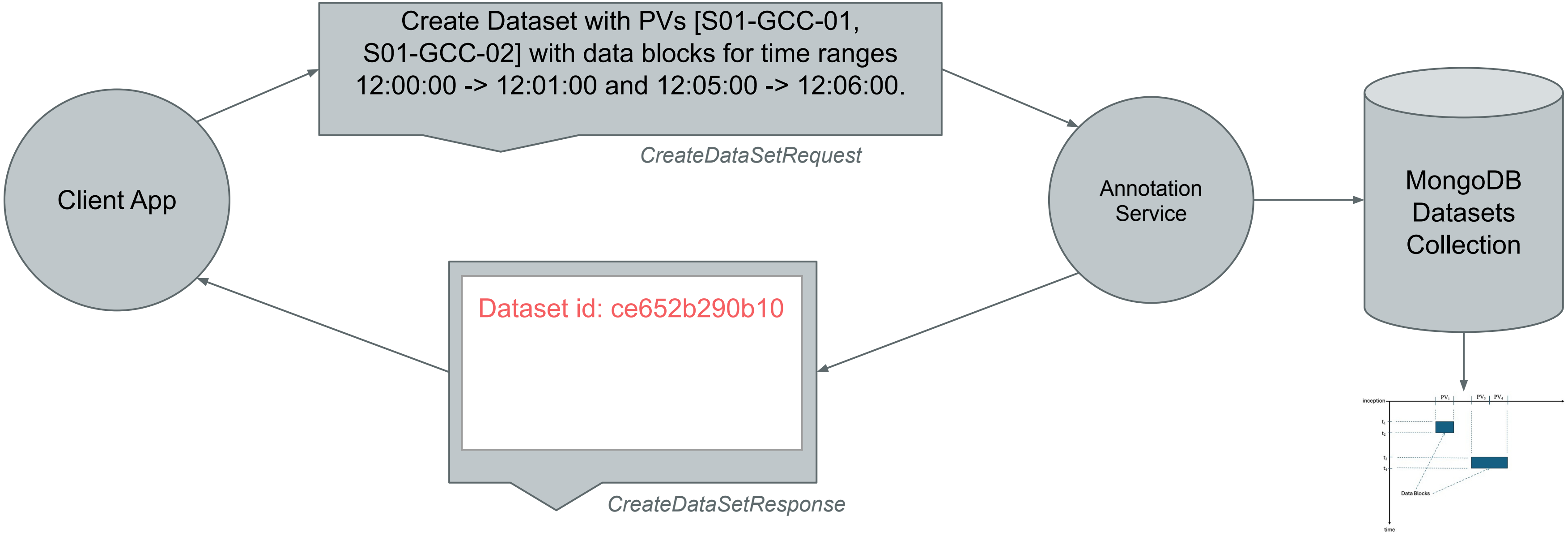
Archive Annotation Datasets

The time-series data archive is analogous to a giant spreadsheet with columns for each PV and rows for each unique timestamp. ***Datasets*** identify subregions of that spreadsheet for targets of the MLDP Annotation and Calculations APIs. A dataset is comprised of ***Data Blocks***, each specifying a list of PV names and a time range.



Archive Annotation Dataset Administration

A client application uses the Annotation Service API to create a Dataset for the specified PV names and time ranges.



Client App

Create Dataset with PVs [S01-GCC-01, S01-GCC-02] with data blocks for time ranges 12:00:00 -> 12:01:00 and 12:05:00 -> 12:06:00.

CreateDataSetRequest

Annotation Service

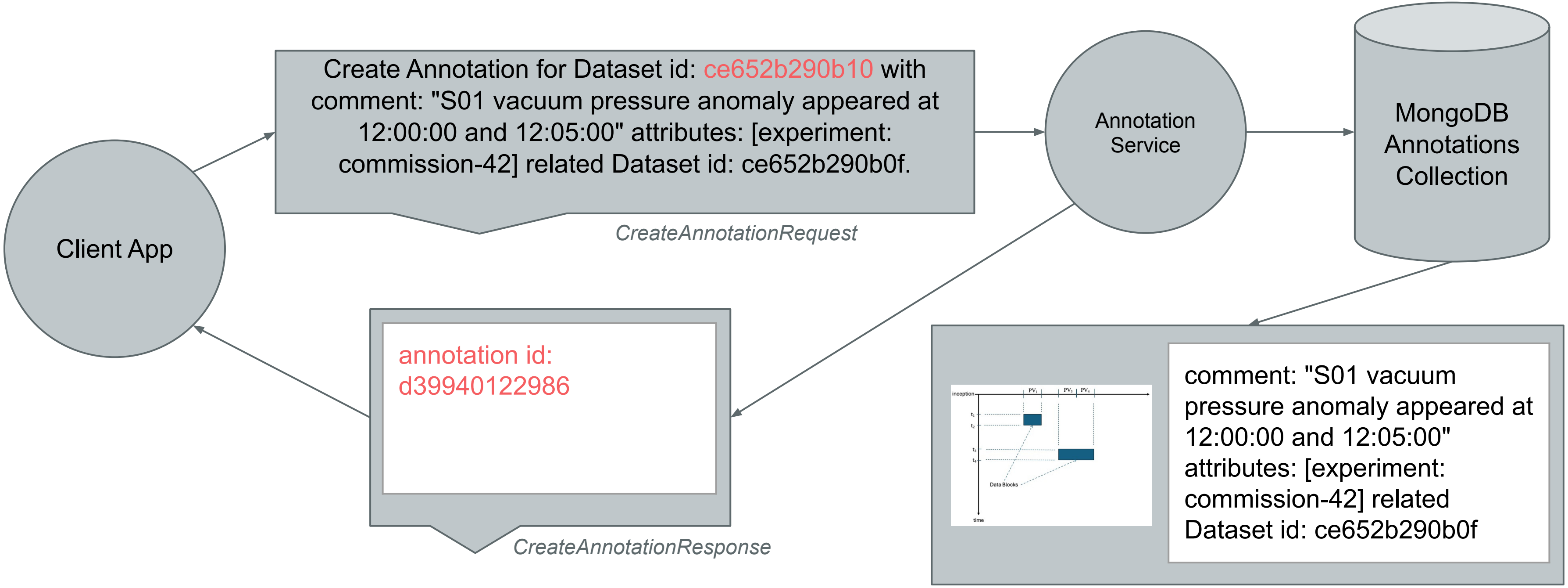
MongoDB Datasets Collection

Dataset id: ce652b290b10

CreateDataSetResponse

Archive Annotation

A client application uses the Annotation Service API to create an Annotation about an anomalous situation for the specified Dataset, specifying the associated experiment name and the id of a related Dataset.



Client App

Create Annotation for Dataset id: ce652b290b10 with comment: "S01 vacuum pressure anomaly appeared at 12:00:00 and 12:05:00" attributes: [experiment: commission-42] related Dataset id: ce652b290b0f.

CreateAnnotationRequest

Annotation Service

MongoDB Annotations Collection

annotation id: d39940122986

CreateAnnotationResponse

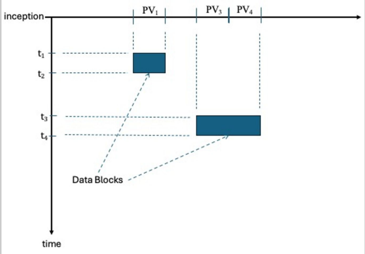
comment: "S01 vacuum pressure anomaly appeared at 12:00:00 and 12:05:00" attributes: [experiment: commission-42] related Dataset id: ce652b290b0f

User-Defined Calculations Handling

A ***Calculations Annotation*** is an Annotation to which Calculations are attached, using a data structure analogous to an Excel workbook with multiple worksheets (the same format used for ingestion of time-series data). For provenance tracking, the Calculations Annotation may reference one or more related Datasets, identifying the data used to derive the Calculations (e.g., normalization of raw data), and include a comment describing the derivation.

comment: The attached Calculations provide normalized values for the S01 vacuum pressure gauge readings over the time range for the linked Dataset.

linked Dataset with raw PV values



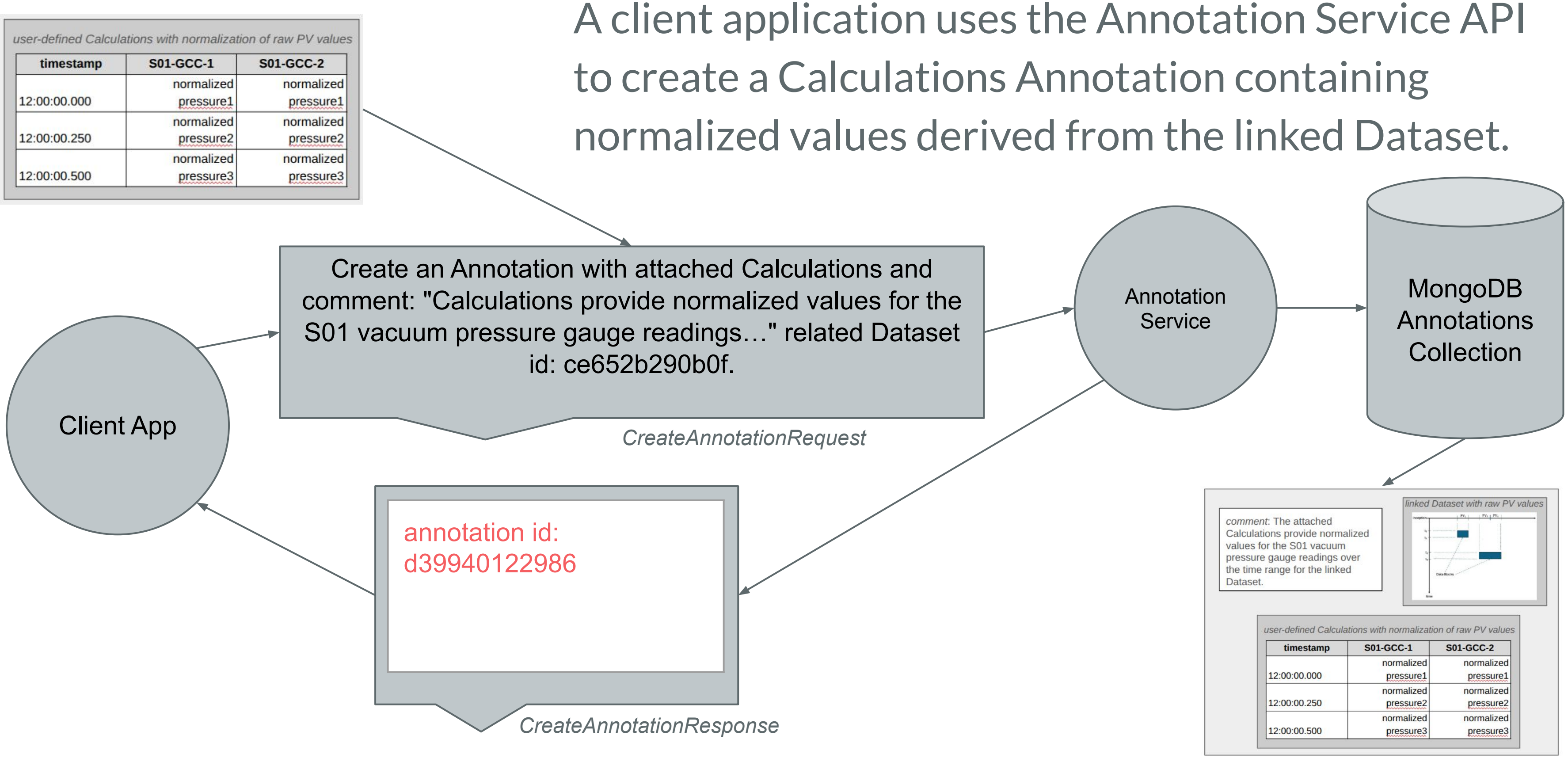
user-defined Calculations with normalization of raw PV values

timestamp	S01-GCC-1	S01-GCC-2
12:00:00.000	normalized pressure1	normalized pressure1
12:00:00.250	normalized pressure2	normalized pressure2
12:00:00.500	normalized pressure3	normalized pressure3

Calculations Annotation

Calculations Upload and Provenance Tracking

A client application uses the Annotation Service API to create a Calculations Annotation containing normalized values derived from the linked Dataset.



user-defined Calculations with normalization of raw PV values

timestamp	S01-GCC-1	S01-GCC-2
12:00:00.000	normalized pressure1	normalized pressure1
12:00:00.250	normalized pressure2	normalized pressure2
12:00:00.500	normalized pressure3	normalized pressure3

Create an Annotation with attached Calculations and comment: "Calculations provide normalized values for the S01 vacuum pressure gauge readings…" related Dataset id: ce652b290b0f.

CreateAnnotationRequest

annotation id: d39940122986

CreateAnnotationResponse

comment: The attached Calculations provide normalized values for the S01 vacuum pressure gauge readings over the time range for the linked Dataset.

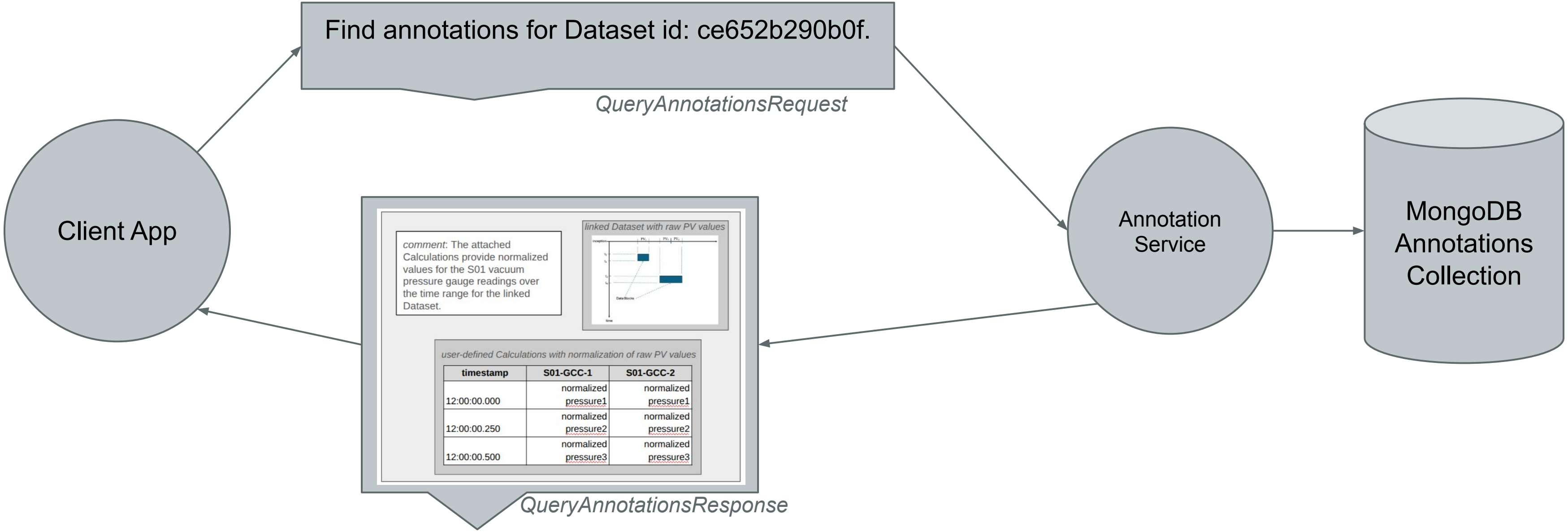
linked Dataset with raw PV values

user-defined Calculations with normalization of raw PV values

timestamp	S01-GCC-1	S01-GCC-2
12:00:00.000	normalized pressure1	normalized pressure1
12:00:00.250	normalized pressure2	normalized pressure2
12:00:00.500	normalized pressure3	normalized pressure3

Archive Annotation Query

A client application searches for Annotations matching query criteria including text content, related Datasets and Annotations, tags, key/value attributes. The response includes details for all matching Annotations including Calculations.



Client App
Find annotations for Dataset id: ce652b290b0f.
QueryAnnotationsRequest
Annotation Service
MongoDB Annotations Collection
linked Dataset with raw PV values
QueryAnnotationsResponse

comment: The attached Calculations provide normalized values for the S01 vacuum pressure gauge readings over the time range for the linked Dataset.

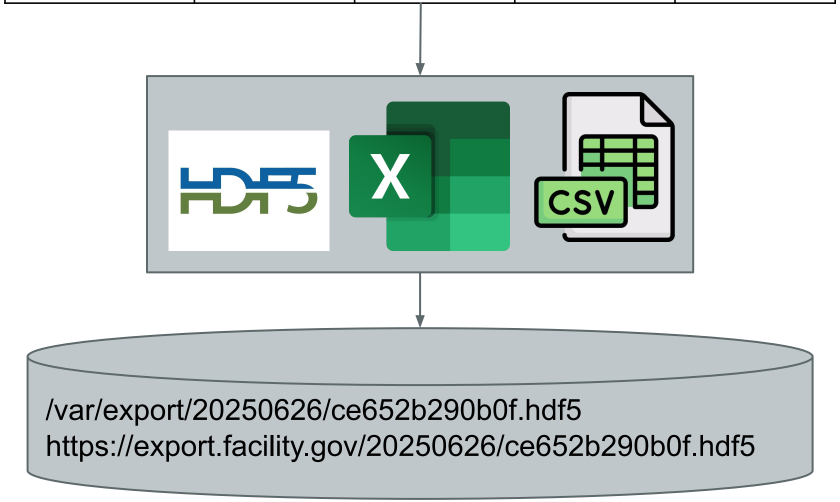
user-defined Calculations with normalization of raw PV values

timestamp	S01-GCC-1	S01-GCC-2
12:00:00.000	normalized pressure1	normalized pressure1
12:00:00.250	normalized pressure2	normalized pressure2
12:00:00.500	normalized pressure3	normalized pressure3

Data and Calculations Export

The export API allows Datasets with raw PV data and user-defined Calculations to be exported to HDF5, XLSX, and CSV files. When both are included in the same export request, the output file includes Calculations data alongside raw PV data.

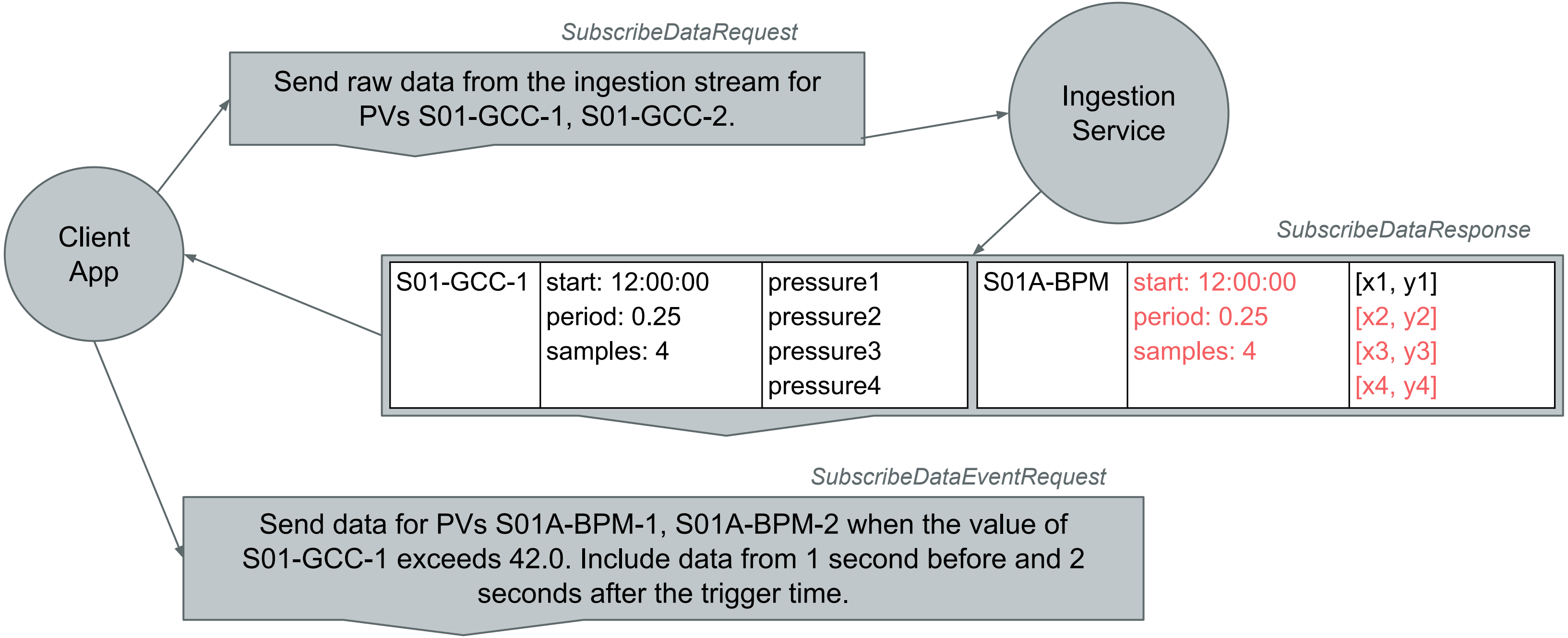
timestamp	S01-GCC-1	S01-GCC-2	normalized gcc-1	normalized gcc-2
12:00:00.000	pressure1	pressure1	normalized pressure1	normalized pressure1
12:00:00.250	pressure2	pressure2	normalized pressure2	normalized pressure2
12:00:00.500	pressure3	pressure3	normalized pressure3	normalized pressure3



/var/export/20250626/ce652b290b0f.hdf5
https://export.facility.gov/20250626/ce652b290b0f.hdf5

Ingestion Stream Subscription

The subscription APIs allow clients to receive raw data for specified PVs or notifications of ***data events*** from the active ingestion stream. Data event subscribers are informed when any of the PV conditions in the data event request are triggered, and can opt to receive data for a list of related target PV names over the time window specified by offset and duration from the trigger time.



SubscribeDataRequest

Send raw data from the ingestion stream for PVs S01-GCC-1, S01-GCC-2.

SubscribeDataResponse

S01-GCC-1	start: 12:00:00 period: 0.25 samples: 4	pressure1 pressure2 pressure3 pressure4	S01A-BPM	start: 12:00:00 period: 0.25 samples: 4	[x1, y1] [x2, y2] [x3, y3] [x4, y4]

SubscribeDataEventRequest

Send data for PVs S01A-BPM-1, S01A-BPM-2 when the value of S01-GCC-1 exceeds 42.0. Include data from 1 second before and 2 seconds after the trigger time.

Ways to Use the MLDP

- low-level gRPC APIs - gRPC support is provided for most programming languages
    - provider registration, query, metadata
    - PV time-series data ingestion, query, subscription
    - ingestion status and error monitoring
    - data set creation, query
    - annotation and calculations creation, query
    - data set and calculations export
    - ingestion stream PV and event subscription
- EPICS aggregator - MLDP data provider for correlated time-series data from EPICS
- Java client libraries - hide low-level gRPC API details
- Java application frameworks - config-driven mechanism for building apps that use the client libraries
- web application
    - navigate data and metadata
    - create data sets
    - annotate archive data
    - export archive data

Learn More

The MLDP is an open-source project, hosted on github at https://github.com/osprey-dcs/data-platform. There you will find complete project documentation with links to repositories for the various components and an installer to get a system up and running quickly.

Osprey

Distributed Control Systems