Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

# ARM's Data Integration Framework

## How ARM (Atmospheric Radiation Measurement) Focuses Its Resources on Science, Not Data Pre-processing

KRISTA GAUSTAD

GABE GIBLER

ARM Infrastructure

https://github.com/ARM-DOE/ADI

https://github.com/ARM-DOE/adi-macosx

Oct 2, 2023

# ARM Data Challenge: Diverse Products Collected and Processed Continuously

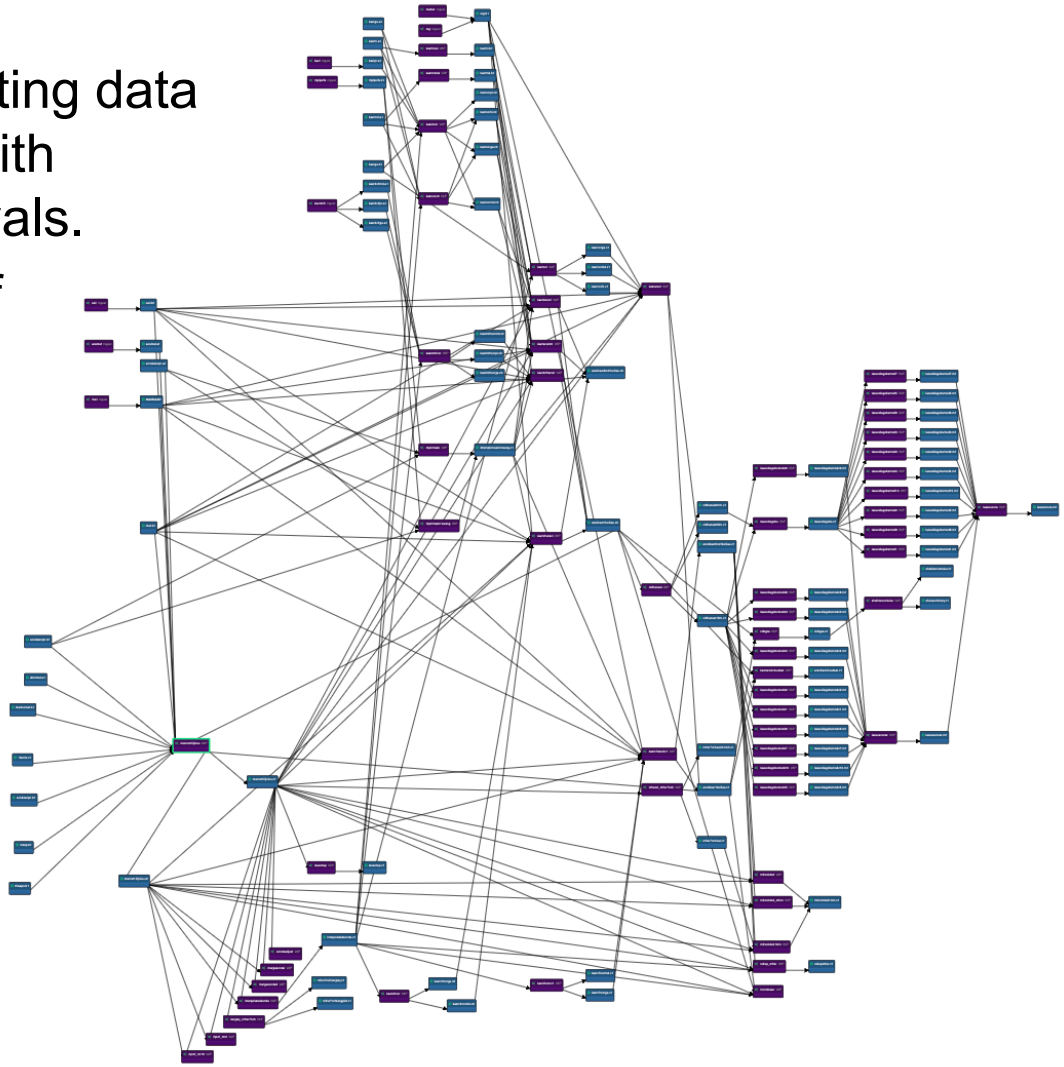
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- 3 fixed sites
- 3 mobile facilities deployed at over 36 locations
- Aerial deployments
- 460+ instruments of approx. 100 types
- 250+ higher level algorithms (VAPs)
- 200–2000 Mb data per day
- 20k–40k files processed per day
- Spanning 30+ years

# ARM Data Challenge: Processing Complex, Interdependent Data

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Ease process of integrating data from diverse datasets with different sampling intervals.
- Simplify development of scientific algorithms.
- Ensure adherence to data standards.



# ARM Data Integrator (ADI)

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965


Location(s)

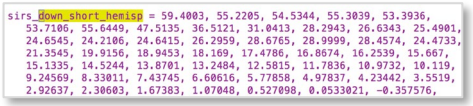
Input Data

Output Data

## PCM (Process Configuration Manager) User Interface

- User-selected measurements with diverse sampling intervals
- User-selected transformation method and parameters

Magic happens…


**Consolidated Data on User-defined Grid**

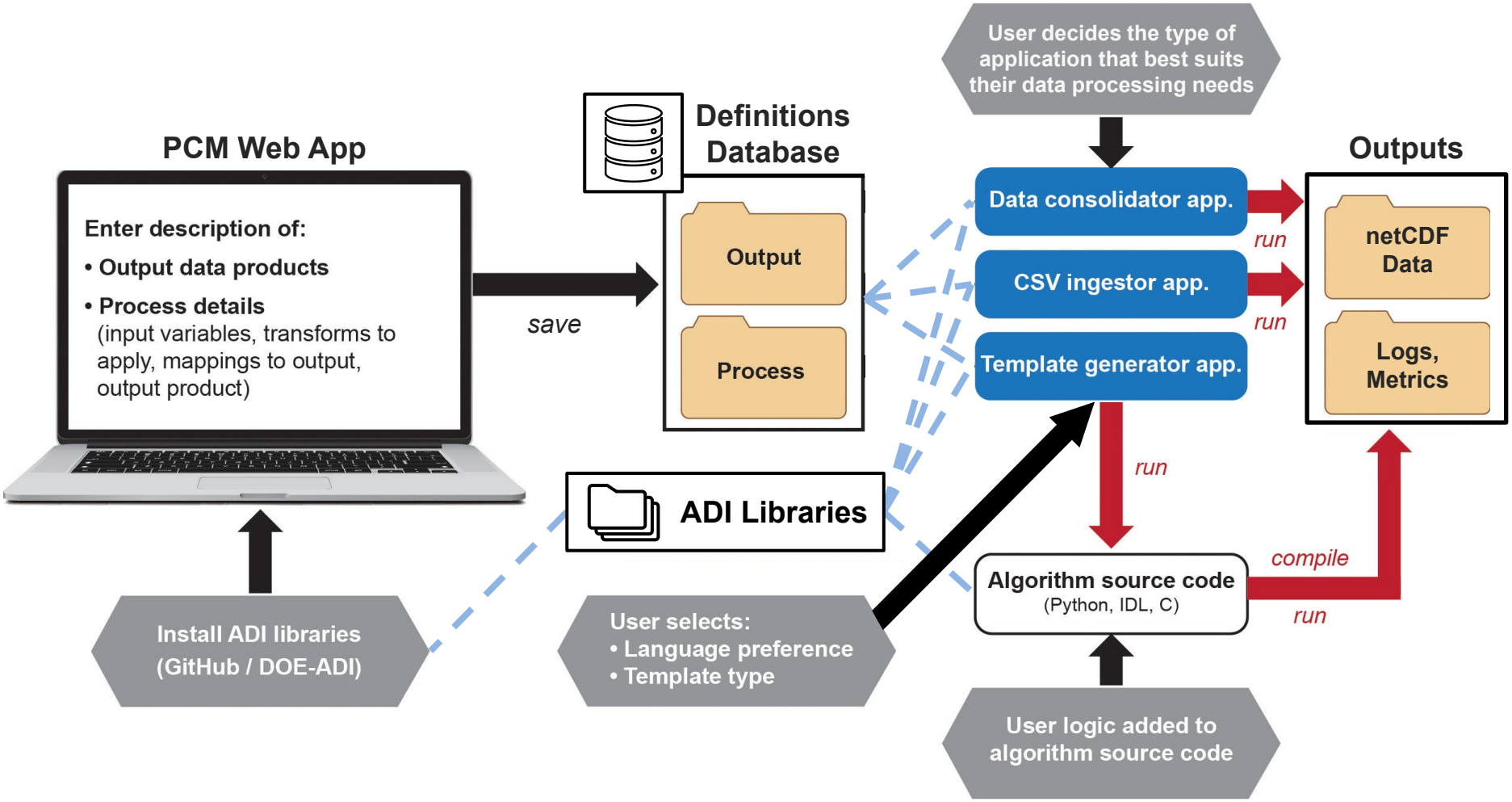
**Python or C Software Project with User Hooks**

# ADI Goals



- Automate repetitive data tasks:
  - Reading input data
  - Transforming and consolidating data
  - Writing data to output
  - Applying or generating quality control flags
- Embed data standards into data design interface.
- Reduce cost, complexity of ingest and scientific algorithm development.
- Improve software robustness and reliability.
- Support multiple programming languages (Python, C).
- Support local installations and open-source code.

# ADI Architecture



# PCM Process: Overview

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ARM PCM Process Manager

Logout (Krista Gaustad) Contact

Home | doe_data_days

**Process: doe_data_days**

Mode: Standard

1: Describe Process

2: Define Inputs & Outputs
- Overview
- Inputs
  - Variables
  - Retrieval Rules
  - Input QC
- Coordinate Systems
  - half_min_grid: time
  - half_min_grid: range
  - metb1: time
- Outputs
  - Output Datastreams
  - Output QC

3: Run Data Consolidator

4: Create ADI Project

Revision History

**Retrieved Variables (4)** View: Interdependencies — Add Variables

Filter by name or partial name

Showing 4 of 4:

**ceil** [ceil.b1]
- **backscatter** (time, range)

**temp30s** [1twrmr.c1, met.b1 ...]
- **temperature** (time)

**tingest** [met.b1]
- **temperature_ingest** (time

**tvap** [1twrmr.c1, tps.b1]
- **temperature_vap** (time)

**Coordinate Systems (2)**
- half_min_grid (time, range)
- metb1 (time)
- No Transformation

**Output Datastreams (2)**
- tutorial.c1
- tutorial30s.c1
- No output datastream defined

# PCM Process: Retrieval Rules

Pacific Northwest NATIONAL LABORATORY
*Proudly Operated by* Battelle *Since 1965*

**Process: doe_data_days** | Mode: Standard

- 1: Describe Process
- 2: Define Inputs & Outputs
  - Overview
  - Inputs
    - Variables
    - Retrieval Rules
    - Input QC
  - Coordinate Systems
    - half_min_grid: time
    - half_min_grid: range
    - metb1: time
  - Outputs
    - Output Datastreams
    - Output QC

**Retrieval Rule Sets (4)** + Add Rule Set

| Name | Rules | Variable Name in Input Datastream |
|---|---|---|
| ceil | | |

| Priority | If process runs at: | Between: | And: | Retrieve data from: |
|---|---|---|---|---|
| 1 | SGP E13 | Any time | Any time | SGP C1 → ceil.b1 |
| 2 | Any sites/facilities | Any time | Any time | Same location as command line → ceil.b1 |

Offsets: Start: 0 seconds End: 0 seconds

| Assigned Name | ceil.b1 (priority 1) | ceil.b1 (priority 2) |
|---|---|---|
| backscatter | backscatter | backscatter |

- Preferred and alternate data sources by:
  - Datastream
  - Location
  - Time
- User-defined variable names can be assigned to the inputs.

# PCM Process: Data Consolidation

Pacific Northwest NATIONAL LABORATORY
*Proudly Operated by* Battelle *Since 1965*

**Process: doe_data_days** | Mode: Standard

- 1: Describe Process
- 2: Define Inputs & Outputs
  - Overview
  - Inputs
    - Variables
    - Retrieval Rules
    - Input QC
  - Coordinate Systems
    - half_min_grid: time
    - half_min_grid: range
    - metb1: time
  - Outputs
    - Output Datastreams
    - Output QC
- 3: Run Data Consolidator
- 4: Create ADI Project
- Revision History

**Dimension Shape**

**How should time coordinate data be determined?** Regular Interval

*time values will be defined using the following parameters:*

**Interval:** 30
**Start:** 0
**End:** 86370
**Number of intervals:** 2880

**Transformation**

**What algorithm should be used to transform variable data dimensioned by time?**

- (●) Auto (average/interpolate)
- ( ) Bin Averaging
- ( ) Bilinear Interpolation
- ( ) Nearest Neighbor (i.e., subsample)
- ( ) Passthrough (i.e., no change in values)

Do you want to override this value by variable? ( ) Yes (●) No

**Output data will be moved to center bin alignment. Do you want to change this setting?** (●) Yes ( ) No

Bin Alignment: Start

**Do you want to adjust the output bin width?** ( ) Yes (●) No

**When reading the input data, how far should the interpolator look for the next valid data point in the time dimension (i.e., the range)?**

Default range for all applicable datastreams: 300 seconds

**Do you want to override this value by datastream?** ( ) Yes (●) No

**Do you want to override this value by variable?** ( ) Yes (●) No

# PCM DOD: Data Output Design

Pacific Northwest NATIONAL LABORATORY
*Proudly Operated by* Battelle *Since 1965*

ARM PCM DOD Manager + New Import Logout (gaustad) Contact Us Help

DODs

tutorial

Filters

Advanced Search

tutorial

Showing 25 of 4955:

- ectutorial30s.c1
- jwmtutorial.c1
- klgtutorial30s.c1
  - DOD v1.0
- qitutorial.c1
- qitutorial30s.c1
- rkntutorial.c1
- rkntutorial30s.c1
- sltutorial.c1
- sltutorial30s.c1
- tutorial.c1
- tutorial30s.c1
- tutorial30smgd.c1
- tutorialcmh.c1
- tutorialcmh30s.c1
- tutorialldr.c1
- tutorialmgd.c1
- tutorialst.c1
- tutorialtsq.c1
- tutorialzcg.c1
- wigtutorial.c1
- wigtutorial30s.c1
- xctutorial.c1
- xctutorial30s.c1
- zftutorial.c1
- zftutorial30s.c1

klgtutorial30s.c1 v1.0

Mode Submit for Review

**Dimensions (3)**

| Dimension | Length |
|---|---|
| time | Unlimited |
| bound | Value 2 |
| range | Value 801 |

**Variables (13)** Filter Configure visible columns

| | name | type | dimension(s) | long_name | units | standard_name | ancillary_variables |
|---|---|---|---|---|---|---|---|
| 0 | base_time | int | | Base time in Epoch | seconds since 1970-1-1 0:00:00 0:00 | <Att Not Defined> | time_offset |
| 1 | time_offset | double | time | Time offset from base_time | <Set at Runtime> | <Att Not Defined> | base_time |
| 2 | time | double | time | Time offset from midnight | <Set at Runtime> | time | <Att Not Defined> |
| 3 | time_bounds | double | time, bound | Time cell bounds | <Att Not Defined> | <Att Not Defined> | <Att Not Defined> |
| 4 | range | float | range | Distance to the center of the corresponding range bin | m | <Att Not Defined> | <Att Not Defined> |
| 5 | range_bounds | float | range, bound | Coordinate_variable cell bounds | <Att Not Defined> | <Att Not Defined> | <Att Not Defined> |
| 6 | temperature | float | time | Air temperature | K | air_temperature | qc_temperature |
| 7 | qc_temperature | int | time | Quality check results on field: Air temperature | unitless | <Att Not Defined> | <Att Not Defined> |
| 8 | backscatter | float | time, range | Backscatter | 1/(sr*km*10000) | <Att Not Defined> | qc_backscatter |
| 9 | qc_backscatter | int | time, range | Quality check results on field: Backscatter | unitless | <Att Not Defined> | <Att Not Defined> |
| 10 | lat | float | | North latitude | degree_N | latitude | <Att Not Defined> |
| 11 | lon | float | | East longitude | degree_E | longitude | <Att Not Defined> |
| 12 | alt | float | | Altitude above mean sea level | m | altitude | <Att Not Defined> |

**Global Attributes (13)**

| Attribute | Type | Value |
|---|---|---|
| command_line | char | <Set at Runtime> |
| Conventions | char | ARM-1.1 |
| process_version | char | <Set at Runtime> |

# PCM DOD: Data Output Validation

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

ARM PCM DOD Manager + New Import

DODs

tutorial

Filters

Advanced Search

tutorial

Showing 25 of 4955:

- ectutorial30s.c1
- jwmtutorial.c1
- klgtutorial30s.c1
  - DOD v1.0
- qitutorial.c1
- qitutorial30s.c1
- rkntutorial.c1
- rkntutorial30s.c1
- sltutorial.c1
- sltutorial30s.c1
- tutorial.c1
- tutorial30s.c1
- tutorial30smgd.c1
- tutorialcmh.c1
- tutorialcmh30s.c1
- tutorialldr.c1
- tutorialmgd.c1

klgtutorial30s.c1 v1.0

## Validation Report for klgtutorial30s.c1: 1.0

3 errors, 9 warnings, 0 disallow (10 can be auto fixed; 0 are ignored)

| Type | Message | Field |
|---|---|---|
| Warning | The calendar variable attribute is optional if the calendar used is Gregorian. | time.calendar |
| Error | "cell_transform" variable attribute value should generally be null, because it is set automatically by ADI. | temperature.cell_transform<br>backscatter.cell_transform |
| Warning | The variable attributes "valid_min", "valid_max" and "valid_range" are intended to describe physical or mathematical limits. For QC test limits, it is suggested to use different attributes, located on the QC variable instead of on the data variable. Suggested attribute names include: "fail_min", "fail_max", "fail_range", "warn_min", "warn_max", "warn_range". | temperature.valid_min<br>temperature.valid_max |
| Warning | Use of the value "unitless" is deprecated. Use "1" instead. | qc_temperature.units<br>qc_backscatter.units |
| Warning | QC variable must have a long_name of "Quality check results on variable: <data variable's long_name attribute value>". | qc_temperature.long_name<br>qc_backscatter.long_name |
| Warning | It is recommended that all QC variables have a "standard_name" attribute with a value of "quality_flag". | qc_temperature.standard_name<br>qc_backscatter.standard_name |
| Error | QC variables using flag_method "bit" and described at the field level must have a description of "This variable contains bit-packed integer values, where each bit represents a QC test on the data. Non-zero bits indicate the QC condition given in the | qc_temperature.description<br>qc_backscatter.description |

(CSV) (TXT) Autofix All Ignore All Violations Close

9 qc_backscatter int time, range Quality check results on field: Backscatter unitless

# ADI Impacts



- Decreased development time and cost.
  - CSV ingest decreased up to 80% (from weeks to days).
  - VAP algorithm decreased up to 60% (from several to few months).
- Improved reliability through code-reuse.
- Software maintenance simplified.
- The processes that generate our data products are more accessible.
- Production processing and reprocessing streamlined.