Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

Provenance Data Harvest and Scientific Results Reproducibility

BIBI RAJU

Pacific Northwest National Laboratory

ESGF 2017, San Francisco, CA

U.S. DEPARTMENT OF
ENERGY

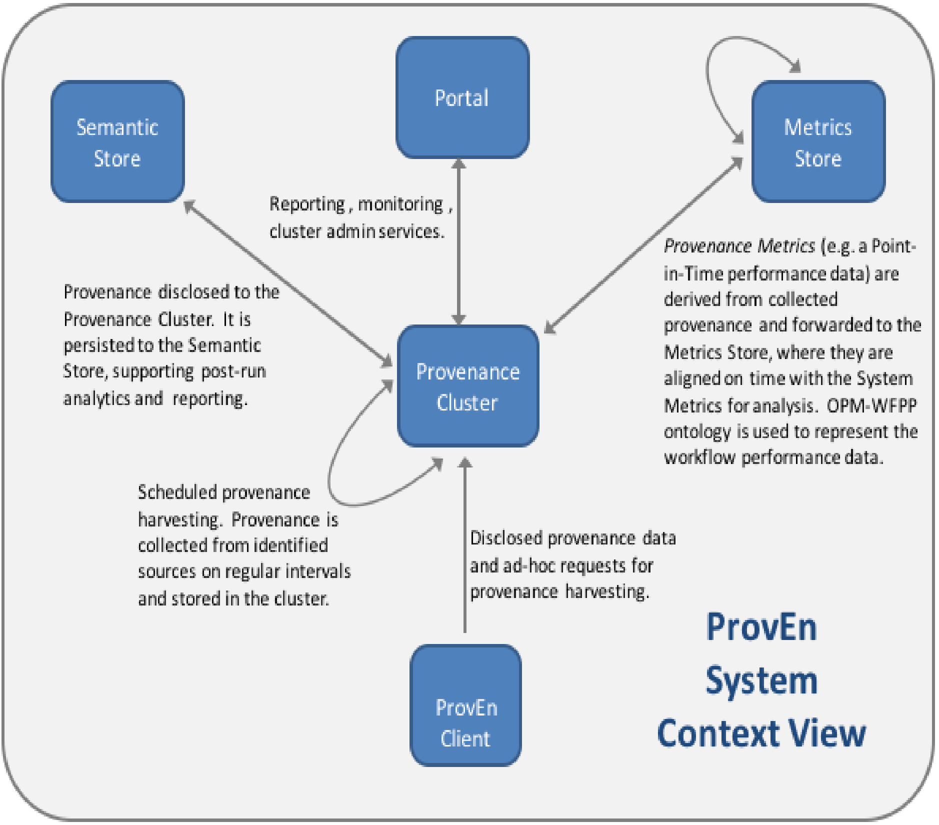
Provenance Environment (ProvEn)

- **ProvEn** is a provenance management platform consisting of loosely coupled components supporting the disclosure, storage, and access to provenance information.

- Describe Anything Provenance Interface API (DAPI)
  - ProvEn's provenance disclosure library. Scientific workflow applications instrumented with DAPI can produced and disclose their provenance data.

- Provenance Cluster
  - ProvEn's scalable approach for collecting concurrent provenance data streams from DAPI sources.

- Hybrid Store
  - ProvEn combines system level metrics (Metric Store) with the traditional disclosed provenance (Semantic Store) to create an extended provenance view.



Reporting , monitoring , cluster admin services.

Provenance disclosed to the Provenance Cluster. It is persisted to the Semantic Store, supporting post-run analytics and reporting.

Provenance Metrics (e.g. a Point-in-Time performance data) are derived from collected provenance and forwarded to the Metrics Store, where they are aligned on time with the System Metrics for analysis. OPM-WFPP ontology is used to represent the workflow performance data.

Scheduled provenance harvesting. Provenance is collected from identified sources on regular intervals and stored in the cluster.

Disclosed provenance data and ad-hoc requests for provenance harvesting.

ProvEn System Context View

ProvEn’s New Features

- HArvester Provenance Interface(HAPI) *New*
  - ProvEn’s harvester library that is capable of extracting already existing file based information produced by applications
  - HAPI uses scruffy provenance content as basis for messages:
    - Tabular data
    - Parameter list
    - Large objects
  - Uses schemas, identifiers, and references to other content to support enrichment.
- Interfaces *New*
  - Developed alpha release portal tied together with Jupyter notebook, Swagger, and REST API and SPARQL endpoint offering a wide range of client side access to provenance

Harvester Provenance Interface (HAPI)

- Extract existing information produced by applications
- Transform the information to HAPI syntax inspired by W3C CSV on the web recommendations
- Pre-stage provenance information into provenance messages
- Write provenance messages into ProvEn store
- Use the retrieved provenance information for
  - scientific results reproducibility
  - scientific results explanations
  - comparing two simulations
- HAPI is a generic format and can be applied to harvest provenance from relational database tables as well as other scientific applications that log provenance related information
- Supports alignment to community vocabularies.
  - Uses W3C PROV for traceability

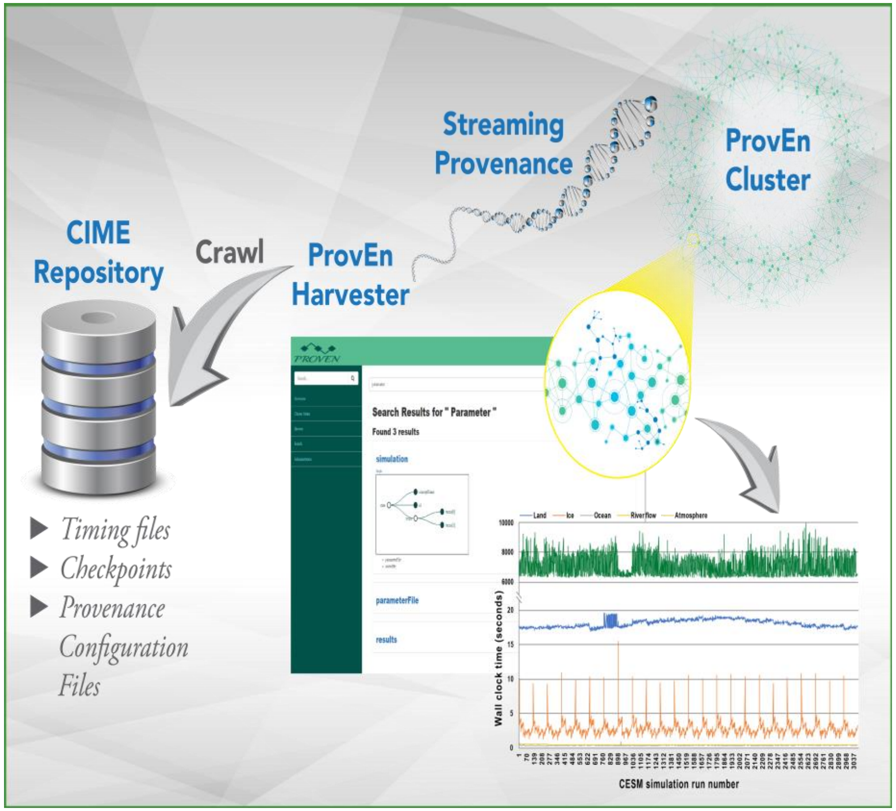
Use Case: Energy Exascale Earth System Model (E3SM)

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

- Focus: Recovering enough information to re-execute a given simulation in the future

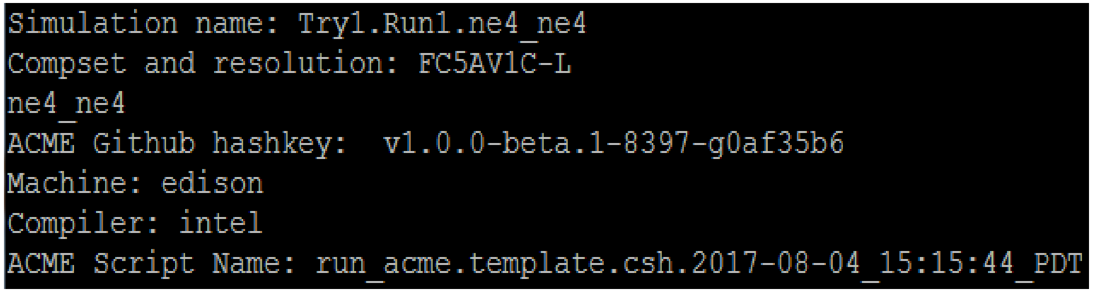
E3SM
Energy Exascale
Earth System Model

- Steps
  - Run a simulation
  - Crawl through simulation data to extract relevant pieces of information
  - Run harvester to store the extracted information in ProvEn database
  - Retrieve captured information from ProvEn database
  - Use the retrieved information to re-execute the simulation with the same set of initial conditions, input parameters and settings on the same machine

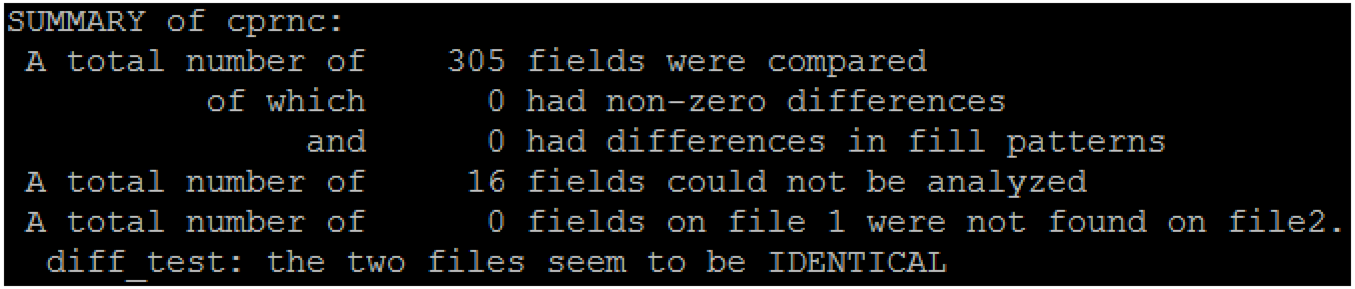


Why is Reproducibility Possible in E3SM?

- E3SM provides a systematic way initializing a directory tree, configuration files, file-based input settings, and run scripts that serve as a base line for any E3SM simulation run
- Simulation code uses configuration control (github) and versioning to manage changes source code, scripts, and new software releases.
- Input files, configuration settings, and scripts were human readable and were easily decipherable.
- Example Artifacts
  - git hash of the E3SM code
  - Machine and compiler details
  - Input parameters
  - simulation compset and resolution
  - configuration XML files



```
Simulation name: Try1.Run1.ne4_ne4
Compset and resolution: FC5AV1C-L
ne4_ne4
ACME Github hashkey:  v1.0.0-beta.1-8397-g0af35b6
Machine: edison
Compiler: intel
ACME Script Name: run_acme.template.csh.2017-08-04_15:15:44_PDT
```



```
SUMMARY of cprnc:
 A total number of    305 fields were compared
          of which      0 had non-zero differences
               and      0 had differences in fill patterns
 A total number of     16 fields could not be analyzed
 A total number of      0 fields on file 1 were not found on file2.
  diff_test: the two files seem to be IDENTICAL
```

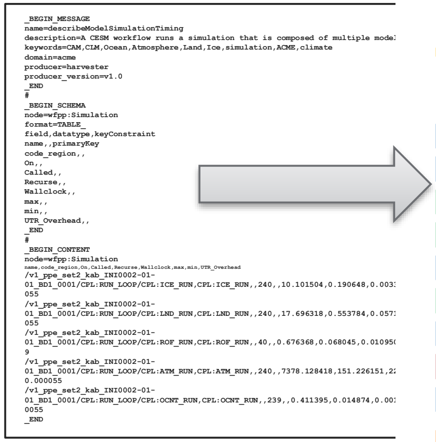
ProvEn Interfaces

- Developed alpha release ProvEn portal that allows visualization of the captured provenance data and Swagger interface for client side access to provenance
- Jupyter notebook interfaced with ProvEn Portal to support desktop analysis
- REST interfaces allow any HTTP enabled client to access time series or semantic information

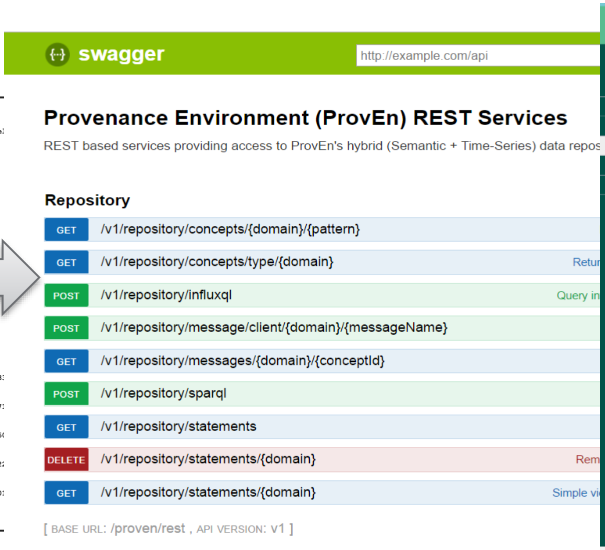
HAPI Message Harvested, Tranformed to JSON-LD

Provenance Fragment saved in named subgraph

Jupyter notebook can be used to query provenance and metrics



```
_BEGIN_MESSAGE
name=describeModelSimulationTiming
description=A CESM workflow runs a simulation that is composed of multiple model
keywords=CAM,CLM,Ocean,Atmosphere,Land,Ice,simulation,ACME,climate
domain=acme
producer=harvester
producer_version=v1.0
_END
#
_BEGIN_SCHEMA
node=wfpp:Simulation
format=TABLE_
field,datatype,keyConstraint
name,,primaryKey
code_region,,
On,,
Called,,
Recurse,,
Wallclock,,
max,,
min,,
UTR_Overhead,,
_END
#
_BEGIN_CONTENT
node=wfpp:Simulation
name,code_region,On,Called,Recurse,Wallclock,max,min,UTR_Overhead
/v1_ppe_set2_kab_INI0002-01-
01_BD1_0001/CPL:RUN_LOOP/CPL:ICE_RUN,CPL:ICE_RUN,,240,,10.101504,0.190648,0.003
055
/v1_ppe_set2_kab_INI0002-01-
01_BD1_0001/CPL:RUN_LOOP/CPL:LND_RUN,CPL:LND_RUN,,240,,17.696318,0.553784,0.057
055
/v1_ppe_set2_kab_INI0002-01-
01_BD1_0001/CPL:RUN_LOOP/CPL:ROF_RUN,CPL:ROF_RUN,,40,,0.676368,0.068045,0.01095
9
/v1_ppe_set2_kab_INI0002-01-
01_BD1_0001/CPL:RUN_LOOP/CPL:ATM_RUN,CPL:ATM_RUN,,240,,7378.128418,151.226151,2
0.000055
/v1_ppe_set2_kab_INI0002-01-
01_BD1_0001/CPL:RUN_LOOP/CPL:OCNT_RUN,CPL:OCNT_RUN,,239,,0.411395,0.014874,0.00
0055
_END
```



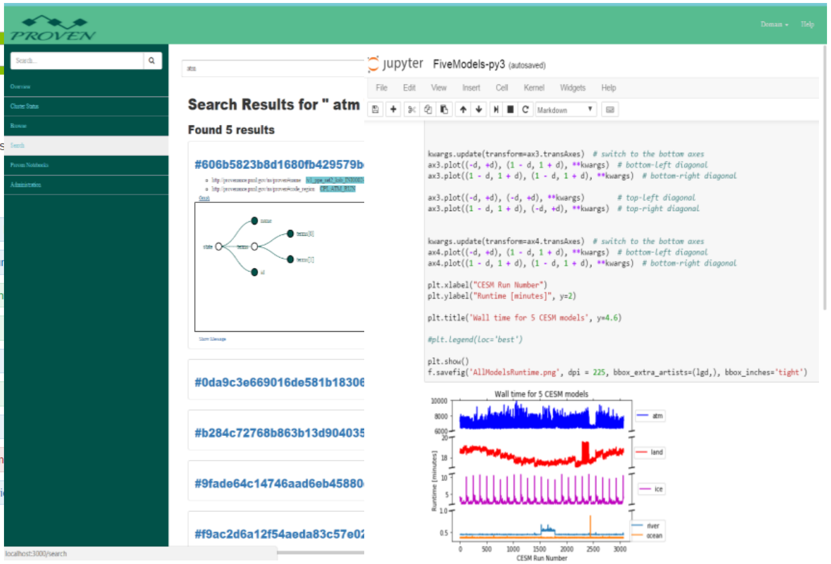
Provenance Environment (ProvEn) REST Services

REST based services providing access to ProvEn's hybrid (Semantic + Time-Series) data repos

Repository

GET	/v1/repository/concepts/{domain}/{pattern}	
GET	/v1/repository/concepts/type/{domain}	Retur
POST	/v1/repository/influxql	Query in
POST	/v1/repository/message/client/{domain}/{messageName}	
GET	/v1/repository/messages/{domain}/{conceptId}	
POST	/v1/repository/sparql	
GET	/v1/repository/statements	
DELETE	/v1/repository/statements/{domain}	Rem
GET	/v1/repository/statements/{domain}	Simple vi

[BASE URL: /proven/rest , API VERSION: v1]



Search Results for " atm

Found 5 results

#606b5823b8d1680fb429579b

- http://provenance.pnnl.gov/ns/proven#name v1_ppe_set2_kab_INI0002-
- http://provenance.pnnl.gov/ns/proven#code_region CPL:ATM_RUN

#0da9c3e669016de581b18306

#b284c72768b863b13d904035

#9fade64c14746aad6eb45880

#f9ac2d6a12f54aeda83c57e02

```
kwargs.update(transform=ax3.transAxes)  # switch to the bottom axes
ax3.plot((-d, +d), (1 - d, 1 + d), **kwargs)  # bottom-left diagonal
ax3.plot((1 - d, 1 + d), (1 - d, 1 + d), **kwargs)  # bottom-right diagonal

ax3.plot((-d, +d), (-d, +d), **kwargs)        # top-left diagonal
ax3.plot((1 - d, 1 + d), (-d, +d), **kwargs)  # top-right diagonal


kwargs.update(transform=ax4.transAxes)  # switch to the bottom axes
ax4.plot((-d, +d), (1 - d, 1 + d), **kwargs)  # bottom-left diagonal
ax4.plot((1 - d, 1 + d), (1 - d, 1 + d), **kwargs)  # bottom-right diagonal

plt.xlabel("CESM Run Number")
plt.ylabel("Runtime [minutes]", y=2)

plt.title('Wall time for 5 CESM models', y=4.6)

#plt.legend(loc='best')

plt.show()
f.savefig('AllModelsRuntime.png', dpi = 225, bbox_extra_artists=(lgd,), bbox_inches='tight')
```

Impact to ESGF

- ProvEn helps in the ESGF domain to maintain
  - detailed history information about the steps followed during data publishing, processing and movement
  - provenance of data products and of the workflows that derive these products and their executions
- Capture provenance in various projects(e.g. CMIP6) for reproducibility
- Extract provenance from projects that already capture provenance
- ProvEn repository could be hosted by those who lack a provenance solution.
- ProvEn is open source (MIT license)

ESGF
Earth System Grid Federation

Acknowledgements

- Eric Stephan, Todd Elsethagen - Pacific Northwest National Laboratory
- Project Acknowledgements
  - Integrated End-to-end Performance Prediction and Diagnosis for Extreme Scientific Workflows (IPPD) Project. IPPD is funded by the U. S. Department of Energy Awards FWP-66406 and DEC0012630
  - Energy Exascale Earth System Model (E3SM) project funded by the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy (DOE) Office of Science.