



U.S. DEPARTMENT OF
ENERGY

Open Energy Data Initiative

Michael Rossol

David Rager

LF Energy Architecture Workgroup

August 19th, 2020



LFENERGY

OEDI

What is OEDI?

The Open Energy Data Initiative (OEDI) aims to improve and automate access of high-value energy data sets across the U.S. Department of Energy's (DOE's) programs, offices, and national laboratories.

Sponsored by DOE, this platform is being implemented by the National Renewable Energy Laboratory (NREL) to make data actionable and discoverable by researchers and industry to accelerate analysis and advance innovation.

Why OEDI?



CLOUD PARTNER
RELATIONSHIPS



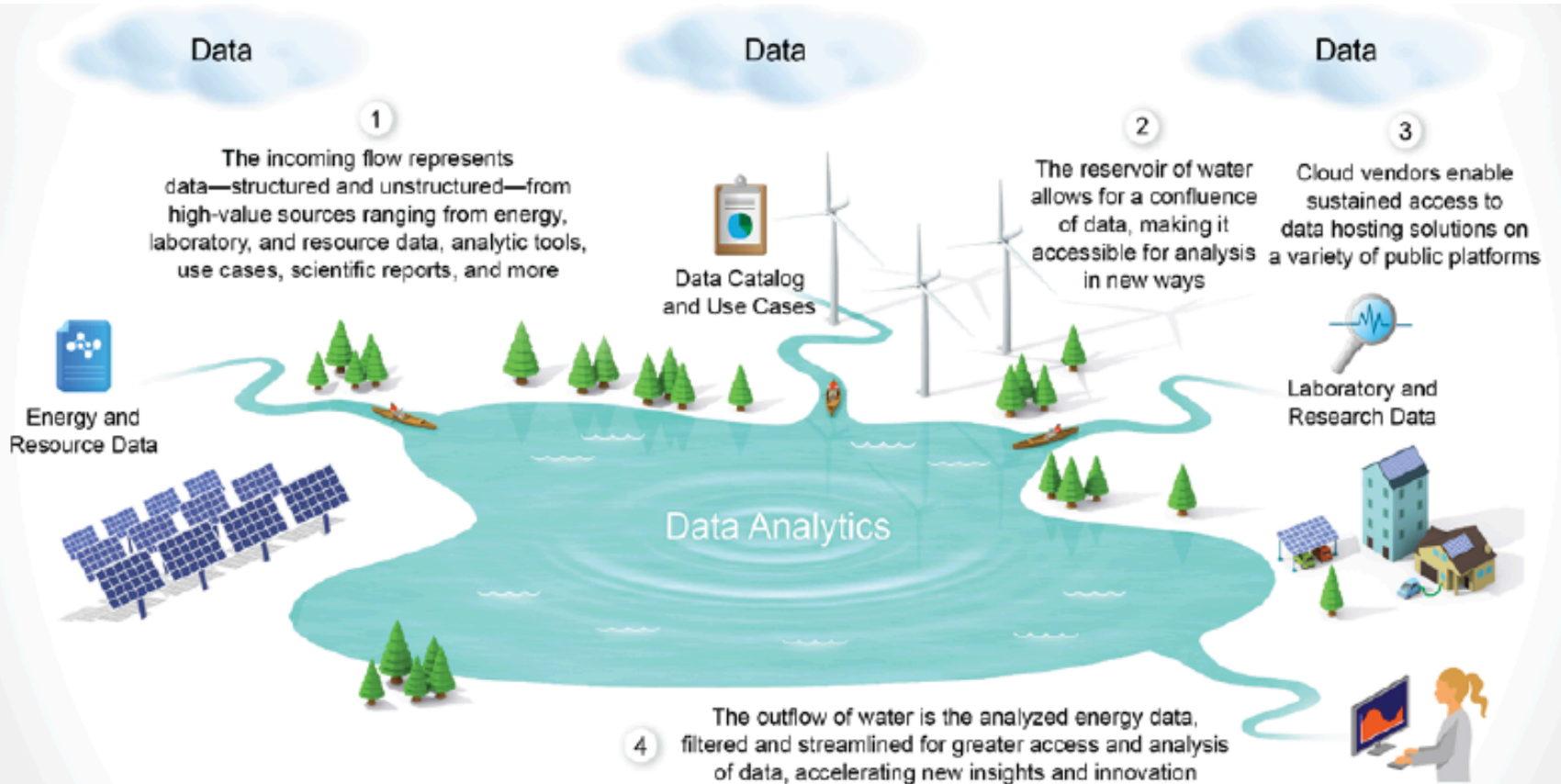
INNOVATIVE
DATASET ACCESS



DATA LAKE &
ANALYTICS

What is OEDI?

All DOE offices + 17 National Labs



OEDI Datasets

Available

National Solar Radiation Database (NDRDB)

Wind Toolkit Dataset

US Historical Wave Data

Porotomo DAS Data

Utility Rate Database

LBL Tracking the Sun Data

PV Rooftop DataBase

In progress

SMART-DS

Cities-LEAP

HTEM Database (High Throughput Experimental Materials)

To Be Linked

GOES NOAA data

US Wind Farm Database (USWFDB)

Geothermal Data Repository (GDR)

Marine Hydro-kinetic Data Repository

Live Wire

Under consideration

PV Module/Cell Performance Data

GHCN-D Climate Dataset

???

What other high value datasets should be included sooner, rather than later?

Amazon Web Services (AWS) Data Lake

Datasets



AWS S3

Cloud Optimized Tools

HSDS



AMAZON ATHENA

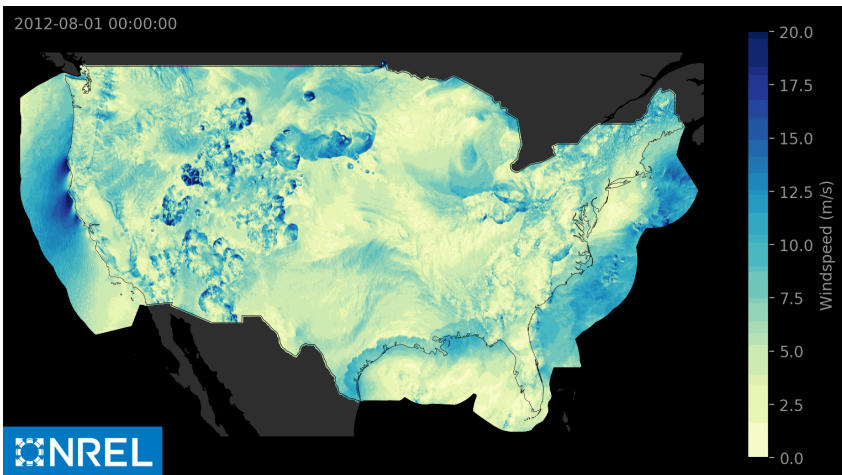
Analysis



Optimizing spatio-temporal data for the cloud



National Solar Radiation Database (NSRDB)

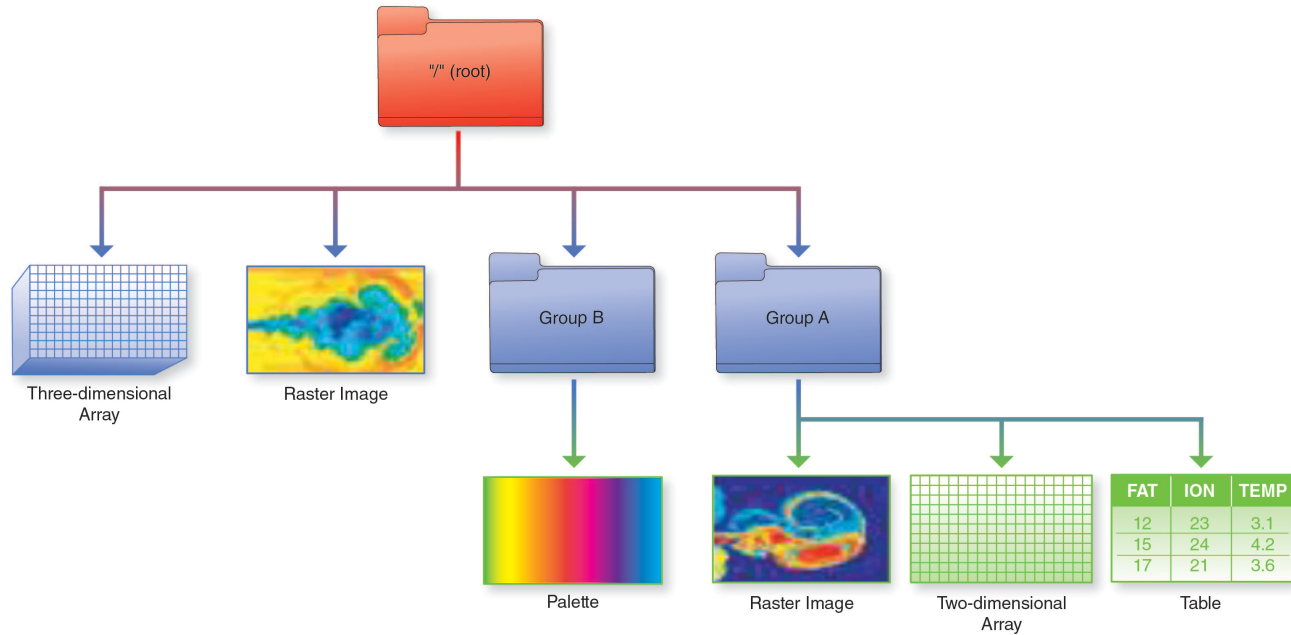


WIND Toolkit

	NSRDB*	WTK
Years	1998-2019	2007-2013
Spatial Resolution	4km x 4km	2km x 2km
Temporal Resolution	30 min	5 min
Geographic Extent	Western Hemisphere	CONUS
File Format	HDF5	HDF5
Size	36 TB	400 TB

*2018 and 2019 5min – 2km is also available

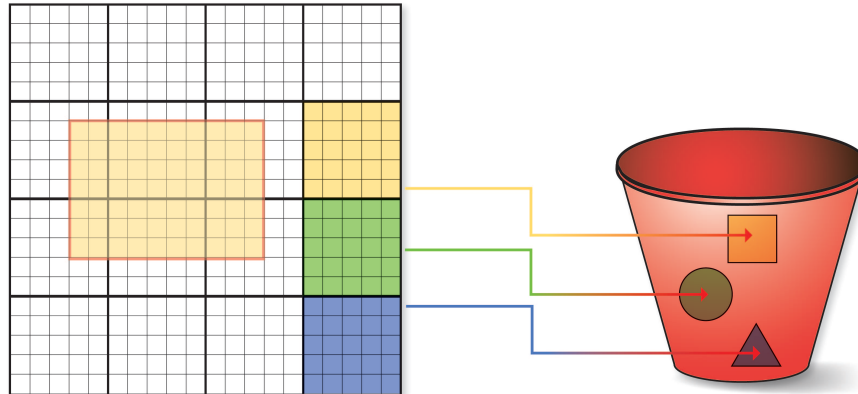
What is HDF?



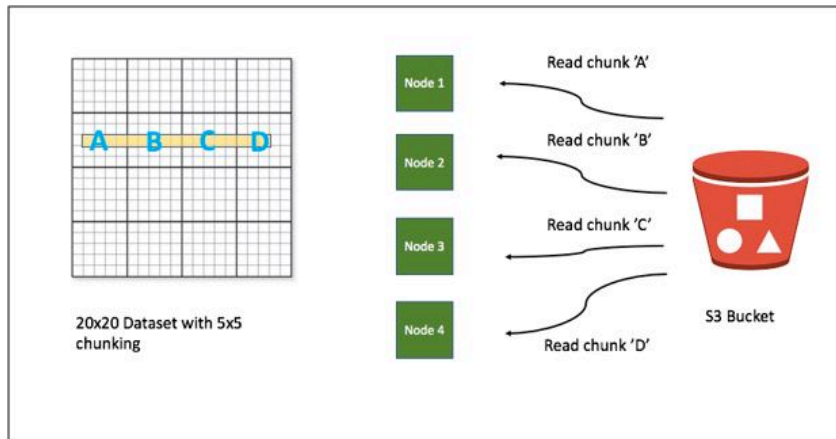
HDF for the cloud -> HSDS

Big Idea: Map individual HDF5 objects (datasets, groups, chunks) as Object Storage Objects

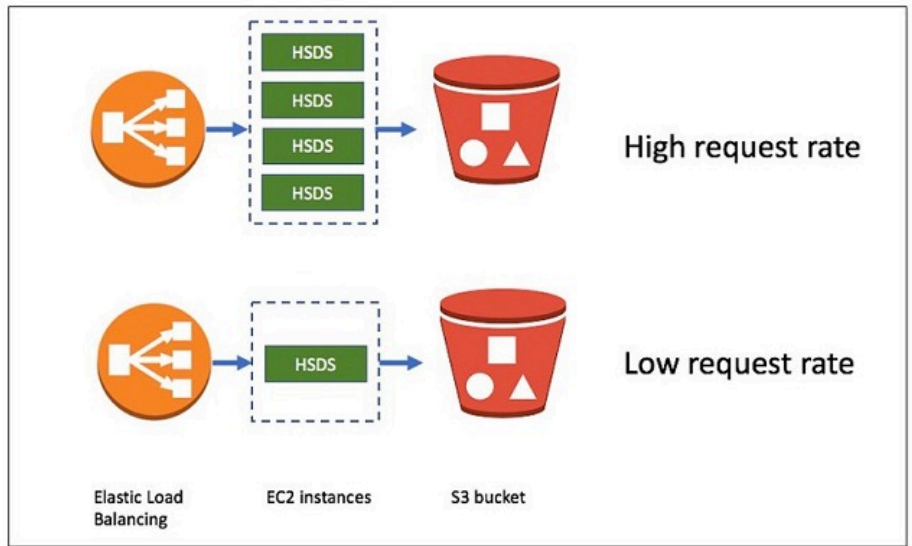
Each chunk (heavy outlines) get persisted as a separate object



Solution: Highly Scalable Data Service (HSDS) HDF + AWS



Parallel requests to S3 allow the HSDS service to scale to the current service demand while not introducing bottlenecks into data flow at the point of data retrieval.
Image Credit: HDF Group.



The HSDS service responds to the request volume by elastically scaling resources.
Image Credit: HDF Group

AWS Athena

API support for multiple languages:



Leverages AWS Services

- Athena – cluster per query
- Glue – simplifies metadata collection and storage

OEDI helps partition the data in S3 for optimal query times

Data Formats Supported

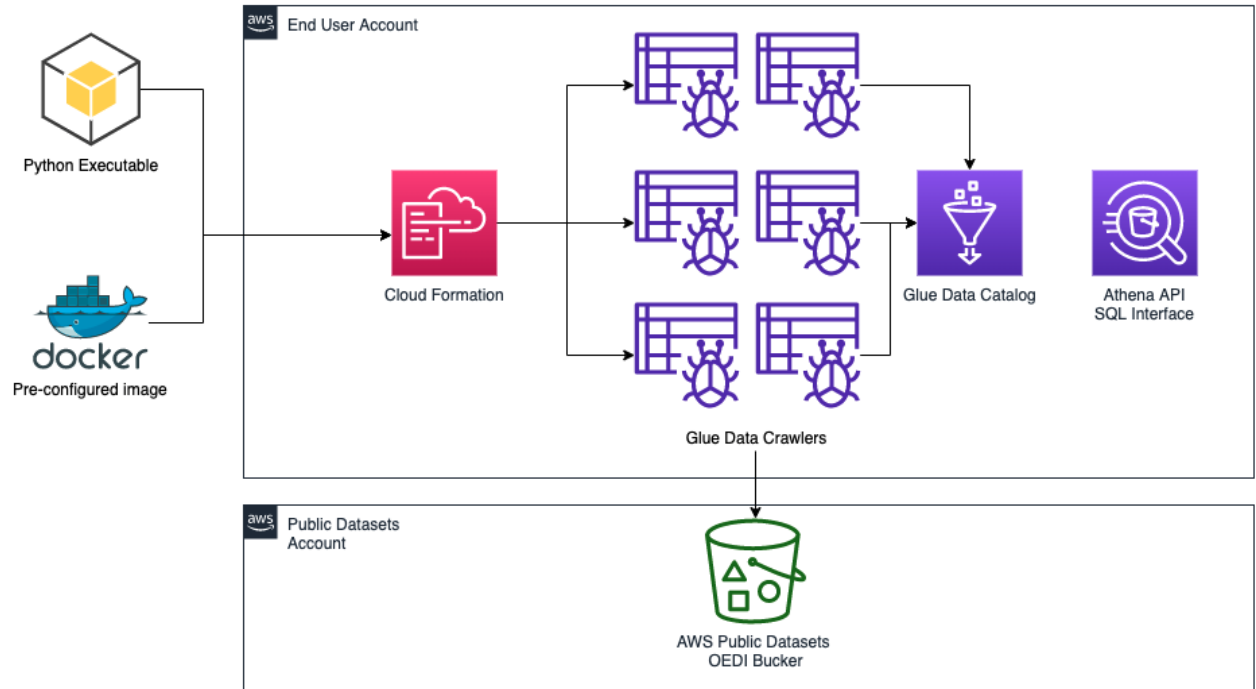
Parquet (preferred), JSON, CSV, TSV, ORC



OEDI Distribution to Enable the AWS APIs:

Python or Docker initiate CloudFormation (infrastructure as code) and the execution of the Glue crawlers.

Once the crawlers create the metadata in Glue Athena is prepared for use with all the included data.



Glue Data Catalog

Table Metadata

Name	nrel_pv_rooftops_aspects												
Description													
Database	oedi_database												
Classification	parquet												
Location	s3://nrel-pv-rooftops/aspects/												
Connection													
Deprecated	No												
Last updated	Wed Jun 10 08:41:39 GMT-600 2020												
Input format	org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat												
Output format	org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat												
Serde serialization lib	org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe												
Serde parameters	<table><tr><td>serialization.format</td><td>1</td></tr></table>	serialization.format	1										
serialization.format	1												
	<table><tr><td>sizeKey</td><td>183832454617</td><td>objectCount</td><td>166</td></tr></table>	sizeKey	183832454617	objectCount	166								
sizeKey	183832454617	objectCount	166										
	<table><tr><td>UPDATED_BY_CRAWLER</td><td>nrel-pv-rooftops-aspects</td></tr></table>	UPDATED_BY_CRAWLER	nrel-pv-rooftops-aspects										
UPDATED_BY_CRAWLER	nrel-pv-rooftops-aspects												
Table properties	<table><tr><td>CrawlerSchemaSerializerVersion</td><td>1.0</td><td>recordCount</td><td>42354990</td></tr><tr><td>averageRecordSize</td><td>404</td><td>CrawlerSchemaDeserializerVersion</td><td></td></tr><tr><td>compressionType</td><td>none</td><td>typeOfData</td><td>file</td></tr></table>	CrawlerSchemaSerializerVersion	1.0	recordCount	42354990	averageRecordSize	404	CrawlerSchemaDeserializerVersion		compressionType	none	typeOfData	file
CrawlerSchemaSerializerVersion	1.0	recordCount	42354990										
averageRecordSize	404	CrawlerSchemaDeserializerVersion											
compressionType	none	typeOfData	file										

Table Schema

	Column name	Data type
1	gid	bigint
2	city	string
3	state	string
4	year	bigint
5	bldg_fid	bigint
6	aspect	bigint
7	the_geom_96703	string
8	the_geom_4326	string
9	region_id	bigint

Partitions

boise_id_07	View files
laguardiajfk_ny_07	View files
tulsa_ok_08	View files
philadelphia_pa_07	View files
dayton_oh_12	View files
worcester_ma_09	View files
huntsville_al_09	View files
providence_ri_04	View files
wichita_ks_12	View files
cleveland_oh_12	View files
greensboro_nc_09	View files
seattle_wa_11	View files

Athena Query Interface

The screenshot displays the Athena Query Interface. On the left, the 'Data source' is set to 'AwsDataCatalog' and the 'Database' is 'oedi_database'. A list of tables is shown, including 'nrel_pv_rooftops_aspects'. The main query editor contains the SQL query: `SELECT * FROM "oedi_database"."nrel_pv_rooftops_aspects" limit 10;`. Below the editor, the 'Run query' button is highlighted, and the status indicates a run time of 4.13 seconds and 93.77 MB of data scanned. The 'Results' section shows a table with 9 rows of data.

Data source [Connect data source](#)

AwsDataCatalog

Database

oedi_database

Filter tables and views...

▼ **Tables (7)** [Create table](#)

- ▶ lbnl_tracking_the_sun_2018 (Partitio... ⋮
- ▶ lbnl_tracking_the_sun_2019 (Partitio... ⋮
- ▼ nrel_pv_rooftops_aspects (Partitio... ⋮
 - gid (bigint)
 - city (string)
 - state (string)
 - year (bigint)
 - bdg_fid (bigint)
 - aspect (bigint)
 - the_geom_96703 (string)
 - the_geom_4326 (string)
 - region_id (bigint)
 - __index_level_0__ (bigint)
 - partition_0 (string) (Partitioned)
- ▶ nrel_pv_rooftops_buildings (Partitio... ⋮
- ▶ nrel_pv_rooftops_developable_plan... ⋮
- ▶ nrel_pv_rooftops_rasd (Partitioned) ⋮
- ▶ nrel_pv_rooftops_syracuse_ny_08... ⋮

New query 1 New query 2 New query 3 **New query 4** +

```
1 | SELECT * FROM "oedi_database"."nrel_pv_rooftops_aspects" limit 10;
```

Run query **Save as** **Create** (Run time: 4.13 seconds, Data scanned: 93.77 MB) **Format query** **Clear**

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

	gid	city	state	year	bdg_fid	aspect	the_geom_96703
1	9959	Pierre	SD	2008	6672	4	MULTIPOLYGON(((−345500.214946738 2384495.65792406,−345502.2090176
2	2250	Pierre	SD	2008	7300	6	MULTIPOLYGON(((−347128.58458669 2385521.58242499,−347128.55515276
3	86036	Pierre	SD	2008	2	4	MULTIPOLYGON(((−346626.427500606 2379788.83920523,−346626.3980286
4	86022	Pierre	SD	2008	2	2	MULTIPOLYGON(((−346622.23344733 2379795.74183314,−346623.23038472
5	86026	Pierre	SD	2008	3	8	MULTIPOLYGON(((−346391.999800867 2379787.01138484,−346392.9967387
6	86027	Pierre	SD	2008	3	2	MULTIPOLYGON(((−346378.072147179 2379785.60104707,−346379.0690850
7	86030	Pierre	SD	2008	3	2	MULTIPOLYGON(((−346377.104686295 2379784.56920724,−346378.1016241
8	86033	Pierre	SD	2008	3	8	MULTIPOLYGON(((−346402.087085709 2379783.2916371,−346402.02813266
9	86034	Pierre	SD	2008	3	8	MULTIPOLYGON(((−346385.139143019 2379782.79668103,−346385.1096661

Questions?

- michael.rossol@nrel.gov
- david.rager@nrel.gov
- data.openei.gov
- github.com/openEDI/documentation
- github.com/nrel/hsds-examples
- registry.opendata.aws