IDRC PEACH DATA HARMONIZATION AND DEPLOYMENT PIPELINE

Contents

Introduction

This report details the pipeline process the PEACH project took in data harmonization for COVID-19 data from source to a common data model, resulting in a harmonized synthetic dataset that was then migrated to a AWS cloud platform for further analysis using a pre-defined study package.

Profiling of data

Two data sets went through the OMOP CDM pipeline that is synthetic COVID-19 data for Malawi and Kenya. Additional data set consisting of a survey done during COVID-19 within Nairobi Kenya urban slum health demographic site was pipelined to the CDM. Data was profiled using white rabbit and additional descriptive in R which both generated excel files with the outputs as below.

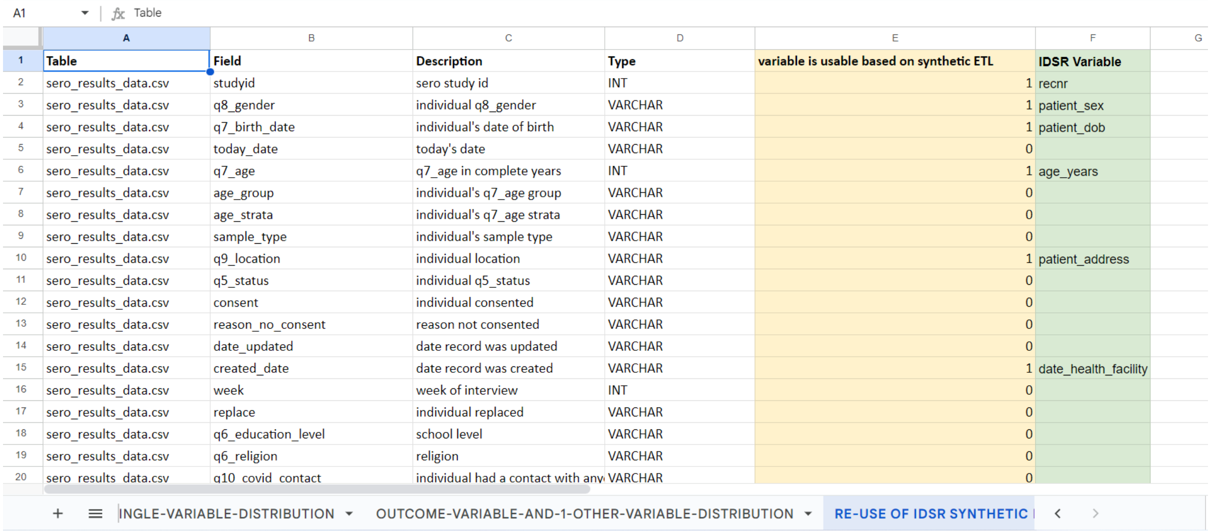

A1
Table
+
INGLE-VARIABLE-DISTRIBUTION
OUTCOME-VARIABLE-AND-1-OTHER-VARIABLE-DISTRIBUTION
RE-USE OF IDSR SYNTHETIC I

Table	Field	Description	Type	variable is usable based on synthetic ETL	IDSR Variable
sero_results_data.csv	studyid	sero study id	INT	1	recnr
sero_results_data.csv	q8_gender	individual q8_gender	VARCHAR	1	patient_sex
sero_results_data.csv	q7_birth_date	individual's date of birth	VARCHAR	1	patient_dob
sero_results_data.csv	today_date	today's date	VARCHAR	0	
sero_results_data.csv	q7_age	q7_age in complete years	INT	1	age_years
sero_results_data.csv	age_group	individual's q7_age group	VARCHAR	0	
sero_results_data.csv	age_strata	individual's q7_age strata	VARCHAR	0	
sero_results_data.csv	sample_type	individual's sample type	VARCHAR	0	
sero_results_data.csv	q9_location	individual location	VARCHAR	1	patient_address
sero_results_data.csv	q5_status	individual q5_status	VARCHAR	0	
sero_results_data.csv	consent	individual consented	VARCHAR	0	
sero_results_data.csv	reason_no_consent	reason not consented	VARCHAR	0	
sero_results_data.csv	date_updated	date record was updated	VARCHAR	0	
sero_results_data.csv	created_date	date record was created	VARCHAR	1	date_health_facility
sero_results_data.csv	week	week of interview	INT	0	
sero_results_data.csv	replace	individual replaced	VARCHAR	0	
sero_results_data.csv	q6_education_level	school level	VARCHAR	0	
sero_results_data.csv	q6_religion	religion	VARCHAR	0	
sero_results_data.csv	q10_covid_contact	individual had a contact with any	VARCHAR	0	

Fig 1 : Profiling of APHRC COVID-19 survey data within NUHDSS

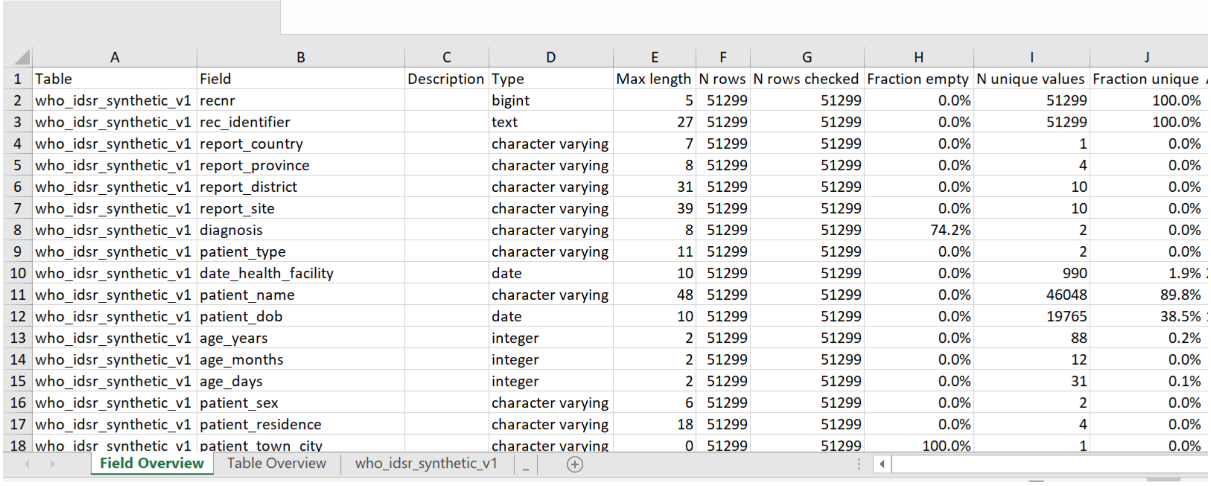


Table	Field	Description	Type	Max length	N rows	N rows checked	Fraction empty	N unique values	Fraction unique
who_idsr_synthetic_v1	recnr		bigint	5	51299	51299	0.0%	51299	100.0%
who_idsr_synthetic_v1	rec_identifier		text	27	51299	51299	0.0%	51299	100.0%
who_idsr_synthetic_v1	report_country		character varying	7	51299	51299	0.0%	1	0.0%
who_idsr_synthetic_v1	report_province		character varying	8	51299	51299	0.0%	4	0.0%
who_idsr_synthetic_v1	report_district		character varying	31	51299	51299	0.0%	10	0.0%
who_idsr_synthetic_v1	report_site		character varying	39	51299	51299	0.0%	10	0.0%
who_idsr_synthetic_v1	diagnosis		character varying	8	51299	51299	74.2%	2	0.0%
who_idsr_synthetic_v1	patient_type		character varying	11	51299	51299	0.0%	2	0.0%
who_idsr_synthetic_v1	date_health_facility		date	10	51299	51299	0.0%	990	1.9%
who_idsr_synthetic_v1	patient_name		character varying	48	51299	51299	0.0%	46048	89.8%
who_idsr_synthetic_v1	patient_dob		date	10	51299	51299	0.0%	19765	38.5%
who_idsr_synthetic_v1	age_years		integer	2	51299	51299	0.0%	88	0.2%
who_idsr_synthetic_v1	age_months		integer	2	51299	51299	0.0%	12	0.0%
who_idsr_synthetic_v1	age_days		integer	2	51299	51299	0.0%	31	0.1%
who_idsr_synthetic_v1	patient_sex		character varying	6	51299	51299	0.0%	2	0.0%
who_idsr_synthetic_v1	patient_residence		character varying	18	51299	51299	0.0%	4	0.0%
who_idsr_synthetic_v1	patient_town_city		character varying	0	51299	51299	100.0%	1	0.0%

Fig 2 : Profiling of IDSR synthetic data

Vocabulary mapping

After the profiling of data, USAGI and Athena were used for vocabulary mapping. The mapped codes were defined clearly in an ETL design document that would be used down the pipeline to guide the work on the actual transformation and loading of data. We relied on OMOP CDM V5 and V6 for these mappings.

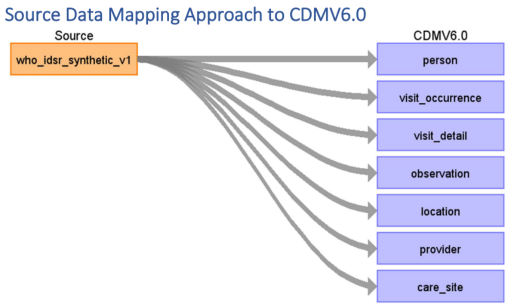
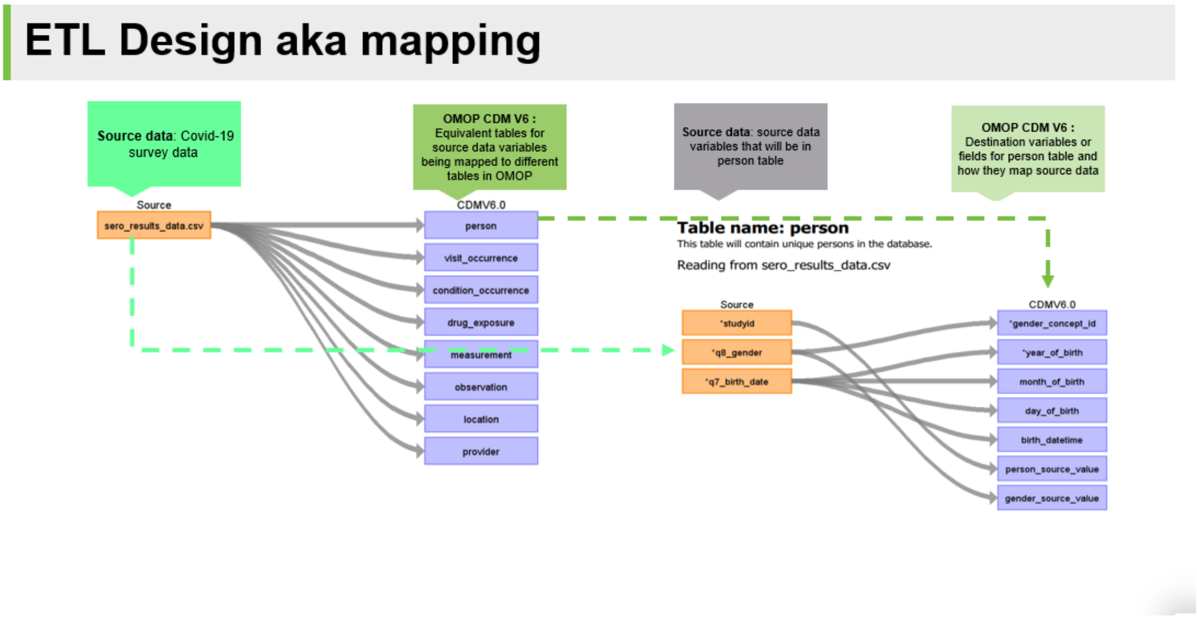


Fig 3: IDSR synthetic data mapping after profiling



ETL Design aka mapping

Source data: Covid-19 survey data

OMOP CDM V6 : Equivalent tables for source data variables being mapped to different tables in OMOP

Source data: source data variables that will be in person table

OMOP CDM V6 : Destination variables or fields for person table and how they map source data

Table name: person

This table will contain unique persons in the database.

Reading from sero_results_data.csv

Fig 4: APHRC COVID-19 survey data mapping

Extract Transform Load Process

ETL followed the mapping stage which generates the ETL design document to help in the ETL development.

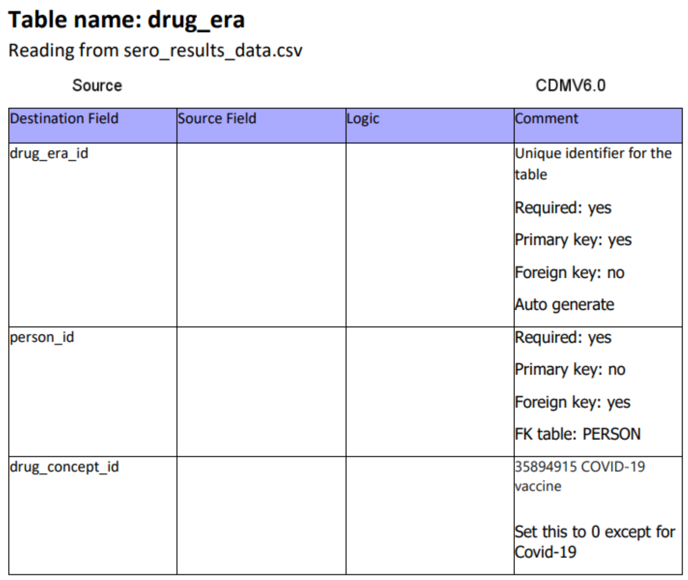


Table name: drug_era

Reading from sero_results_data.csv

Source CDMV6.0

Destination Field	Source Field	Logic	Comment
drug_era_id			Unique identifier for the table Required: yes Primary key: yes Foreign key: no Auto generate
person_id			Required: yes Primary key: no Foreign key: yes FK table: PERSON
drug_concept_id			35894915 COVID-19 vaccine Set this to 0 except for Covid-19

Fig 5 : Sample ETL design document for drug era table

The ETLs were developed using a java tool called Pentaho and SQL query language. Based on the design document scripts were generated to transform source data to the mapped standard concepts supported within an OMOP CDM v5.4. The detail on how each transformation is programmed is bundled within each step and can be interrogated given the need.

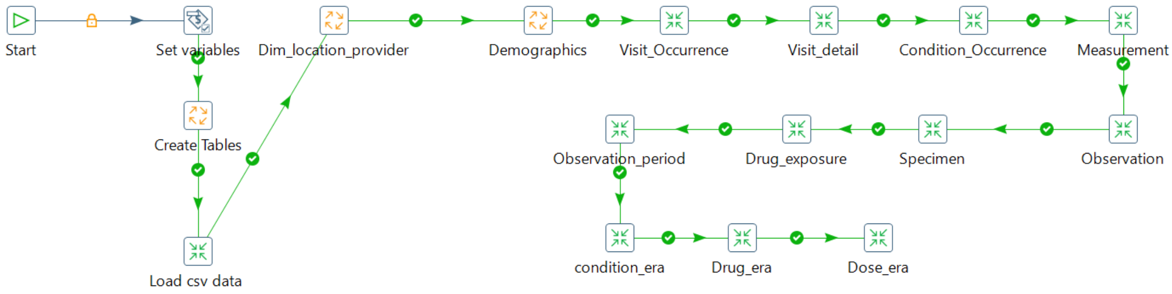


Fig 6 : Pentaho ETL pipeline with various transformations and jobs

The output of each ETL leads to the population of respective OMOP CDM tables. Depending on the data and efficient design this should run pretty fast.

Deployment of OMOP CDM and results schemas to AWS

Given the harmonized data from source as OMOP CDM, the next step is to run an OHDSI tool called Achilles to generate aggregated data called results schema. With both CDM and results schema the deployment stage is set.

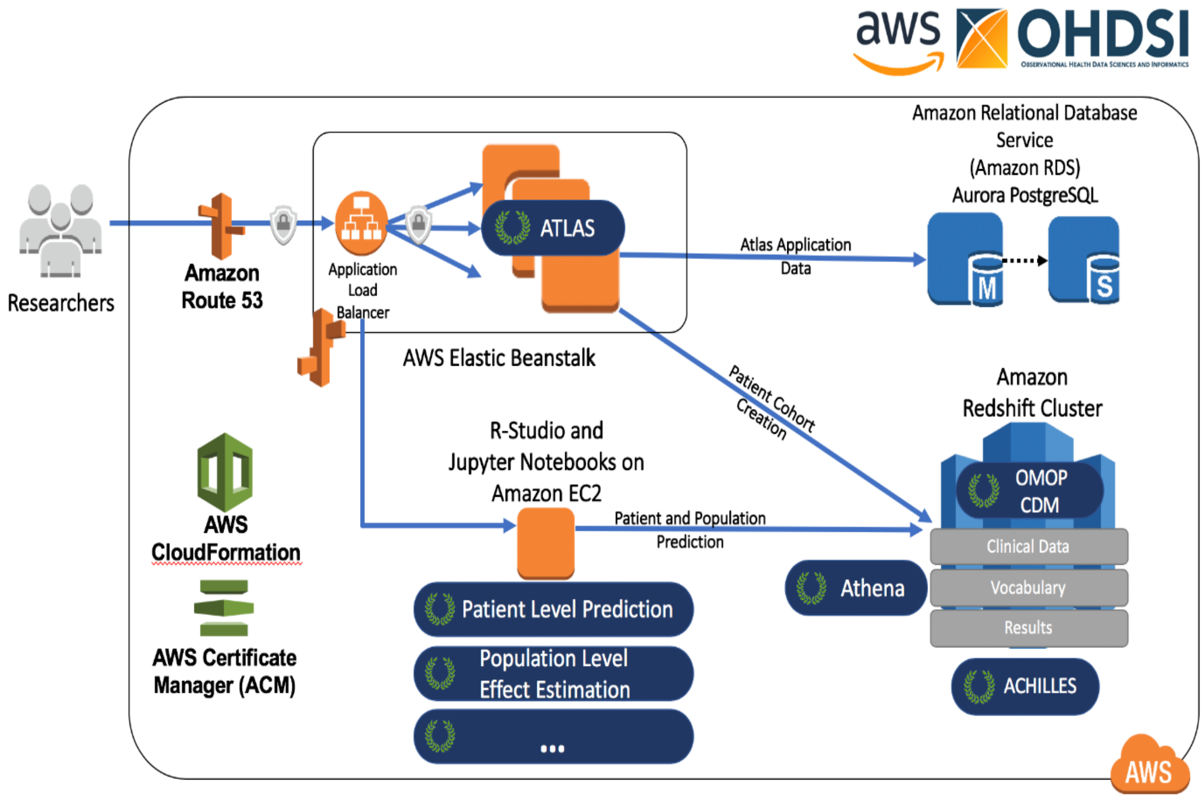

aws
OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS
Researchers
Amazon Route 53
Application Load Balancer
ATLAS
AWS Elastic Beanstalk
Atlas Application Data
Amazon Relational Database Service (Amazon RDS) Aurora PostgreSQL
M
S
Patient Cohort Creation
Amazon Redshift Cluster
OMOP CDM
Clinical Data
Vocabulary
Results
ACHILLES
Athena
R-Studio and Jupyter Notebooks on Amazon EC2
Patient and Population Prediction
Patient Level Prediction
Population Level Effect Estimation
...
AWS CloudFormation
AWS Certificate Manager (ACM)
AWS

Fig 7: OHDSIonAWS deployment architecture

Outline of process involved

- Create text files for each database table to be pushed to AWS
- Log in to AWS account with both console and programmatic access
- Create schema for the vocabularies, results and CDM
- Create S3 bucket and move data from the local CDM and results schema to S3
- Pull the data from S3 bucket to Redshift

Challenges and lessons learned

We faced a number of challenges during the process of harmonizing peach data worth mentioning for instance:

- Deciding on which ETL tool to use, we wanted a tool that everyone can learn quickly, we settled on SQL and Pentaho although SQL and Python was an alternative. Whatever the choice, the team had to learn either Pentaho or Python. Pentaho seemed easier to learn and we settled on it.
- Learning the various tools chosen within the shortest time possible to a level of being productive with the tool. This included most of the OHDSI, ETL and cloud deployment tools available for OMOP CDM.

- Migrating data to the cloud for the federated analysis was another challenge as there were no open source tools for this, we had to write custom code in python to move million rows of data to cloud (S3 and Redshift)
- Adding our own data set to AWS Redshift OMOP CDM database failed. We sorted help online and it seemed like a bug from AWS side on provisioning this particular OHDSIonAWS stack. We approached the lead architect James for this and he was able to ascertain this was a bug from their deployment limiting addition of new datasets to Atlas. He corrected this and a new stack was provisioned for our use and the greater public good.
- Managing and adding more users to Atlas. The stack only comes provisioned with two users, we currently rely on manual addition of users but this is an area worth spending more resources on.

Conferences attended

INSPIRE AGM, Nairobi Kenya

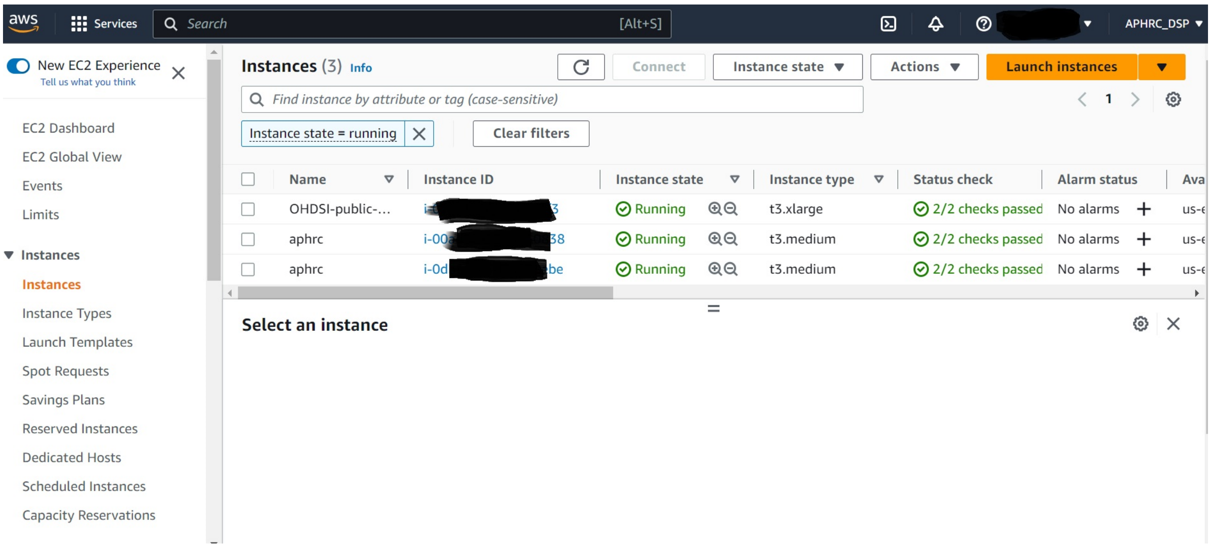
- presentation: Introduction to OMOP CDM
- Date : February 2023
- Link : https://drive.google.com/file/d/1oJOlo7L3pW1VNqK4lk827kAoZL7kVM8K/view?usp=sharing

LAISDAR AGM, Kigali Rwanda

- presentation: Experiences of data harmonization in Kenya. use case PEACH project
- Date : 15-16 June 2023
- Link : https://drive.google.com/file/d/14LvKYHmfr5Buz5Iv2fmS0cz1voUjEN_8/view?usp=sharing

Achievements and project outputs

Data harmonization was successful despite enormous challenges we faced during the implementation phase as evident through various research outputs and testimonials. For the first time in Africa we were able to harmonize data for COVID-19 from different countries with different study settings. Outputs screenshots attached below.

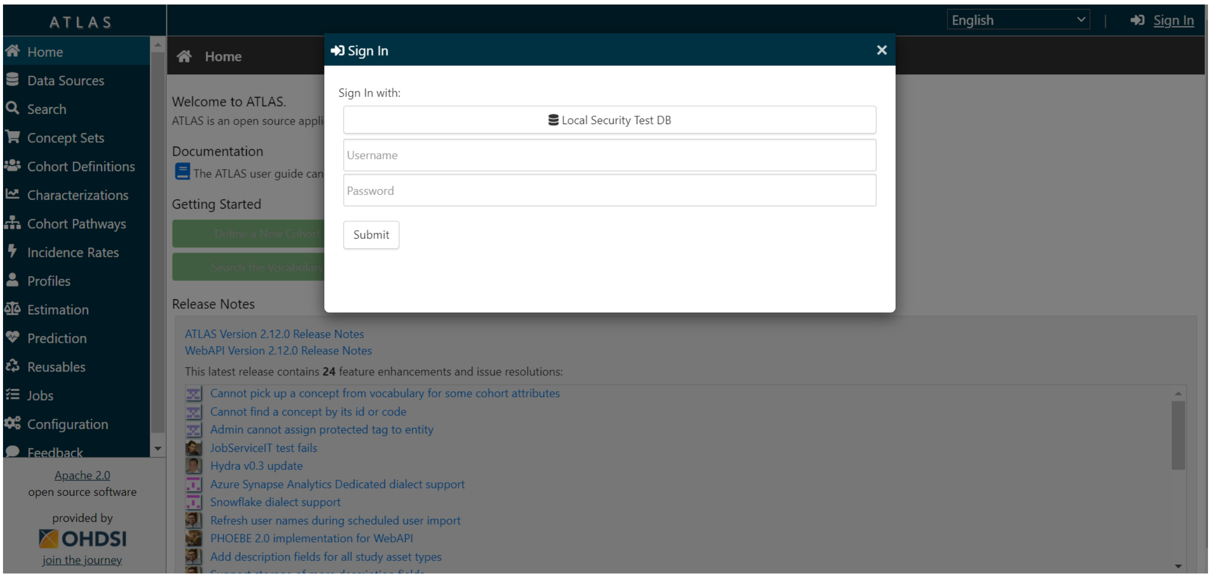


[redacted]

Instance state = running

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Ava
OHDSI-public-...	i[redacted]3	Running	t3.xlarge	2/2 checks passed	No alarms	us-e
aphrc	i-00[redacted]38	Running	t3.medium	2/2 checks passed	No alarms	us-e
aphrc	i-0d[redacted]be	Running	t3.medium	2/2 checks passed	No alarms	us-e

Fig 8 : OHDSI on Aws Instances



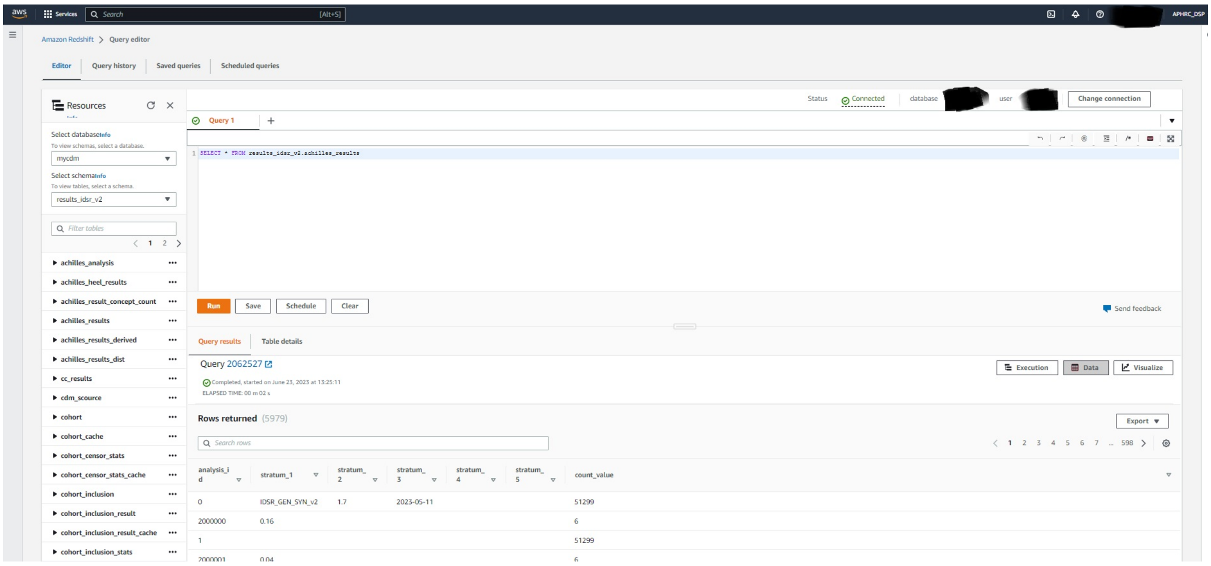
Release Notes

ATLAS Version 2.12.0 Release Notes
WebAPI Version 2.12.0 Release Notes

This latest release contains **24** feature enhancements and issue resolutions:

- Cannot pick up a concept from vocabulary for some cohort attributes
- Cannot find a concept by its id or code
- Admin cannot assign protected tag to entity
- JobServiceIT test fails
- Hydra v0.3 update
- Azure Synapse Analytics Dedicated dialect support
- Snowflake dialect support
- Refresh user names during scheduled user import
- PHOEBE 2.0 implementation for WebAPI
- Add description fields for all study asset types

Fig 9: Atlas login page for aggregated analysis



```
SELECT * FROM results_idsr_v2.achilles_results
```

Query 2062527

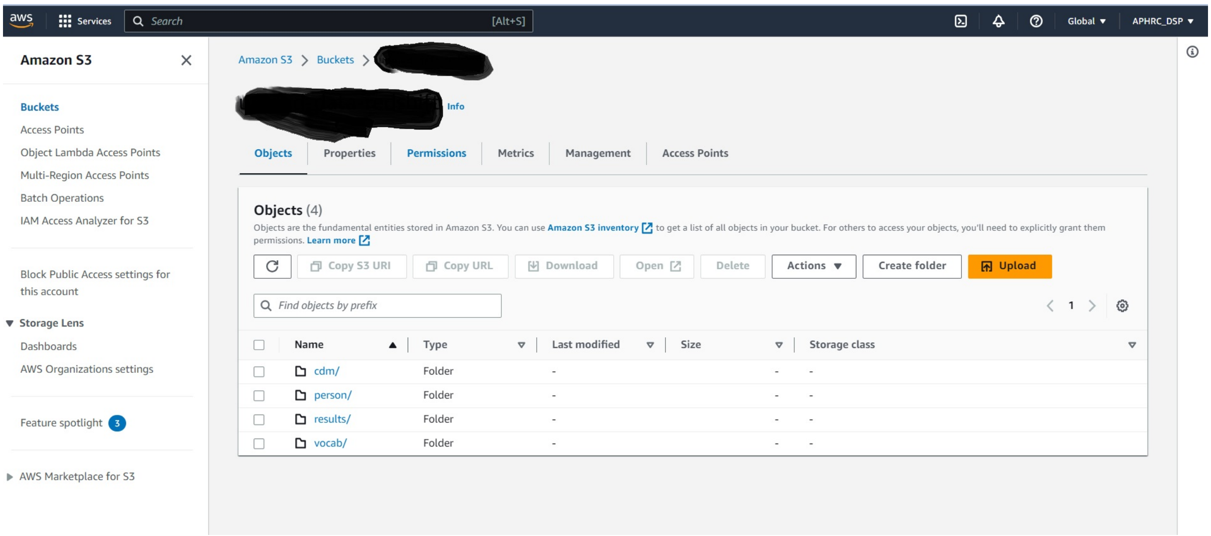
Completed, started on June 23, 2023 at 13:25:11

ELAPSED TIME: 00 m 02 s

Rows returned (5979)

analysis_id	stratum_1	stratum_2	stratum_3	stratum_4	stratum_5	count_value
0	IDSR_GEN_SYN_v2	1.7	2023-05-11			51299
2000000	0.16					6
1						51299
2000001	0.04					6

Fig 10 : A view of Redshift database



Objects (4)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	cdm/	Folder	-	-	-
□	person/	Folder	-	-	-
□	results/	Folder	-	-	-
□	vocab/	Folder	-	-	-

Fig 11 : Simple storage service (S3) for temporarily hosting data during migration process to redshift

Useful Links

I. Data profiling and mapping

II. ETL design

III. AWS deployment set-up

IV. Data migration from on-prem to AWS Redshift via Simple storage service (S3)

V. LAISDAR Report on using OHDSI on-prem.

VI. Atlas Login URL