Прикладне програмування в науках про Землю

доц. Онищук В.І.
доц. Демидов В.К.
аспір. Охрімчук Р.Ю.

2024

Загальні відомості

Open Data Cube (ODC) — це відкрите програмне забезпечення для організації доступу, управління та аналізу великих обсягів геоданих.

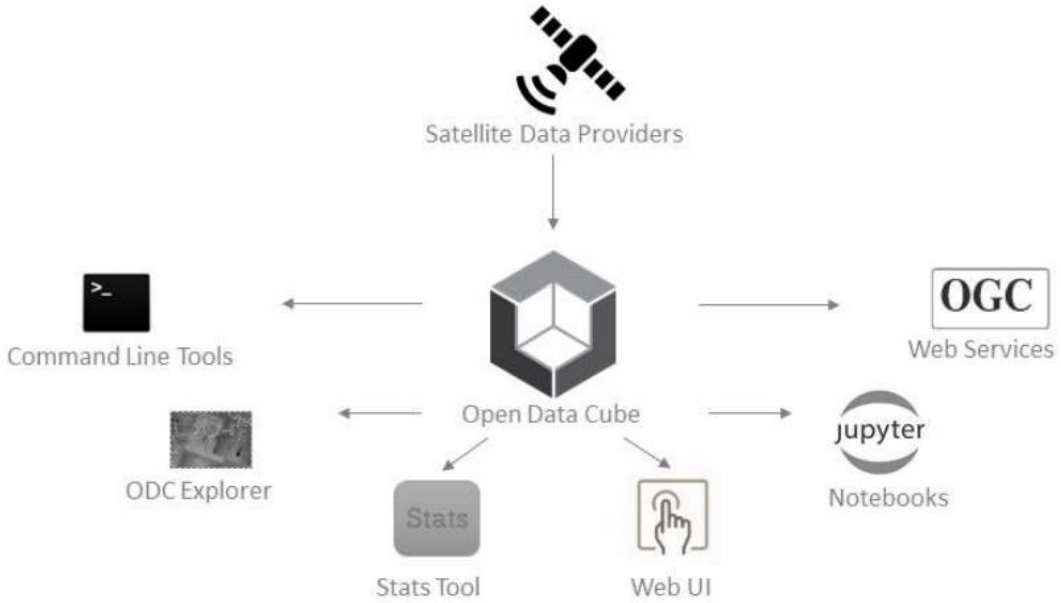


Figure 2: High-Level ODC Ecosystem

Джерело: https://www.opendatacube.org/_files/ugd/3632b4_269d1d61d7f04677a1d32278042aa51a.pdf

Посилання для ознайомлення з потенціалом зазначеної технології:
https://www.youtube.com/watch?v=QQe2YYy9xAU

Приклади впровадження ODC

Розташування осередків розвитку даної технології станом на 2017 рік

Джерело: https://medium.com/opendatacube/what-is-open-data-cube-805af60820d7

Найкращі реалізації ODC на даний момент:

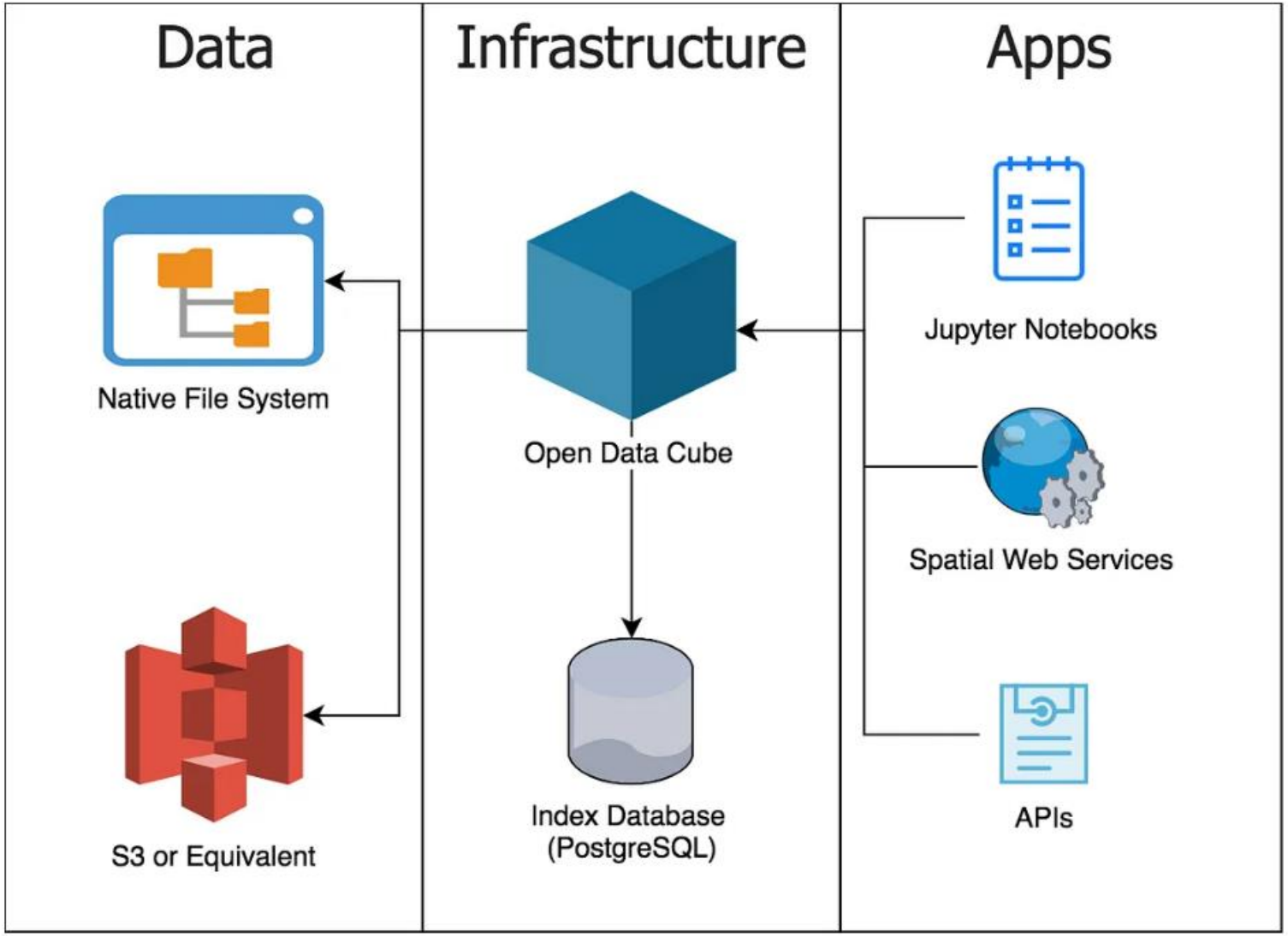
Digital Earth AUSTRALIA

https://docs.dea.ga.gov.au/notebooks/Beginners_guide/README

Digital Earth AFRICA

https://docs.digitalearthafrica.org/en/latest/maps/deafrica_map.html

Спрощення схема будови та взаємозв'язку компонентів ODC

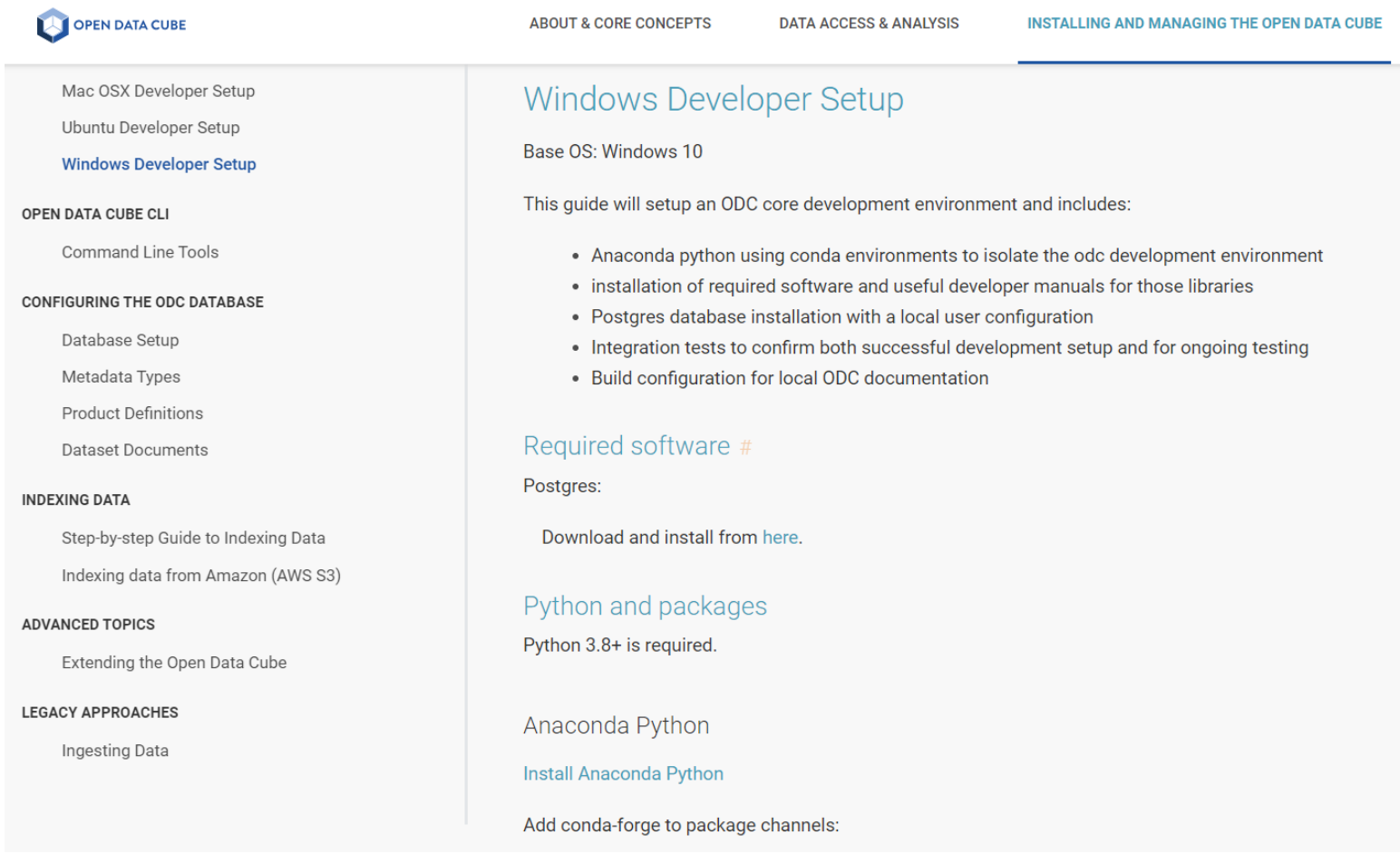


Джерело: https://medium.com/opendatacube/what-is-open-data-cube-805af60820d7

Рекомендації щодо налаштування ODC

Компоненти, які потрібно завантажити та інсталювати, для встановлення ODC:

- *завантажити та налаштувати середовище anaconda, python 3.8+;*
- *встановити бібліотеку datacube (ця операція автоматично також завантажить допоміжні бібілотеки для стабільної роботи ODC;*
- *встановити Postgres 14+*



Посилання на інструкцію інсталяції ODC:
https://opendatacube.readthedocs.io/en/latest/installation/setup/windows.html

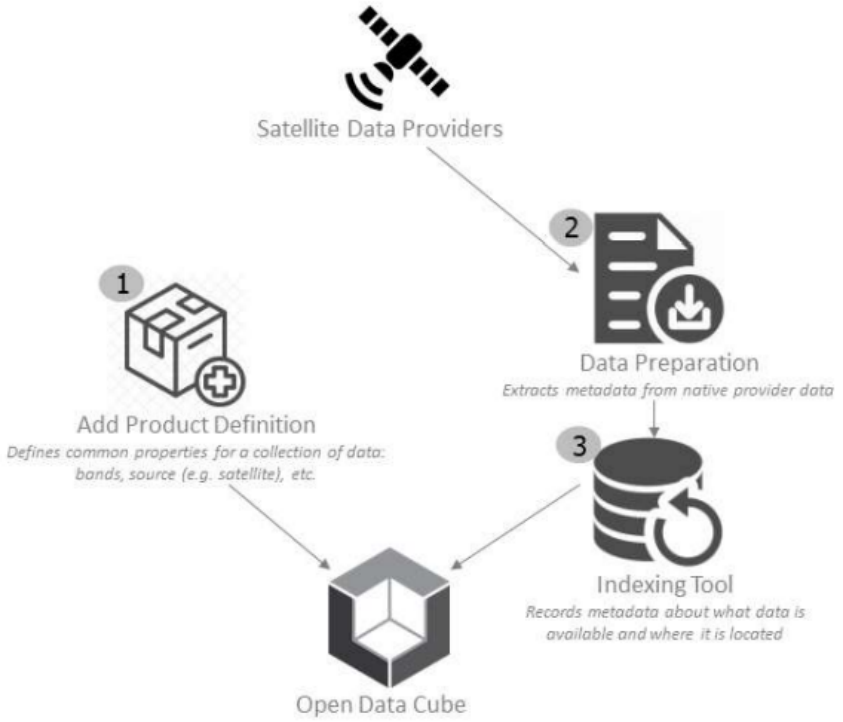
Індексація даних

У даному контексті термін "індексація" визначає процес підготовки та додавання метаданих, які містять інформацію про продукт та датасети, до бази даних. Існують три типи метаданих: eo, telemetry та eo3. Типи eo та telemetry застаріли і застосовуються виключно у рамках вже реалізованих проектів. Формат EO3 є сучасним, і саме з ним ми працюватимемо.

Процес індексації включає два етапи:

1. Створення файлу метаданих, що описує майбутній продукт. На цьому етапі необхідно проявити пильність, оскільки після створення продукту вносити зміни до його структури буде неможливо;

2. Формування файлу метаданих для датасету. На цьому етапі потрібно розробити код, що буде читати або штучно генерувати GeoTransform та CRS, які деталізують просторове положення цільової сцени та визначають шляхи до розташування кожного з файлів, згідно з кількістю measurements, ініційованих при створенні продукту.



Джерело: https://www.opendatacube.org/_files/ugd/3632b4_269d1d61d7f04677a1d32278042aa51a.pdf

Деталі щодо створення продукту

```
---
name: dem_srtm  (назва продукту)
metadata_type: eo3 (тип метаданих)

metadata:
  product:
    name: dem_srtm  (назва продукту)

measurements: (це словник, який місить інформації про всі змінні, які містить даний продукт)
  - name: elevation  (назва змінної)
    dtype: int16 (тип даний)
    nodata: -32768.0  (значення для маркування nodata)
    units: "metre"  (одиниці вимірювання)
```

Технічні деталі за посиланням:
https://opendatacube.readthedocs.io/en/latest/installation/product-definitions.html

Приклади метаданих для Landsat продуктів:
https://docs.digitalearthafrica.org/en/latest/data_specs/Landsat_C2_SR_specs.html

Деталі щодо створення датасетів

```yaml
# UUID of the dataset
id: f884df9b-4458-47fd-a9d2-1a52a2db8a1a
$schema: 'https://schemas.opendatacube.org/dataset' (вказує на приналежність до формату eo3)

# Product name
product:
  name: landsat8_example_product  (назва продукту, яка була вказала у файлі із метаданими продукту при його створені)

# Native CRS, assumed to be the same across all bands
crs: "epsg:32660"  (система координат)

# Optional GeoJSON object in the units of native CRS.
# Defines a polygon such that all valid pixels across all bands
# are inside this polygon.
geometry:
  type: Polygon
  coordinates: [[..]]  (координати полігону, який описує валідне покриття вздовж всіх змінних)

# Mapping name:str -> { shape:     Tuple[ny: int, nx: int]
#                       transform: Tuple[float x 9]}
# Captures image size, and geo-registration
grids:
    default:  # "default" grid must be present
       shape: [7811, 7691]
       transform: [30, 0, 618285, 0, -30, -1642485, 0, 0, 1]
    pan:  # Landsat Panchromatic band is higher res image than other bands
       shape: [15621, 15381]
       transform: [15, 0, 618292.5, 0, -15, -1642492.5, 0, 0, 1]

# Per band storage information and references into `grids`
# Bands using the "default" grid should not need to reference it
measurements:
   pan:               # Band using non-default "pan" grid
     grid: "pan"      # should match the name used in `grids` mapping above
     path: "pan.tif"
   red:               # Band using "default" grid should omit `grid` key
     path: red.tif    # Path relative to the dataset location
   blue:
     path: blue.tif
   multiband_example:
     path: multi_band.tif
     band: 2          # int: 1-based index into multi-band file
   netcdf_example:    # just example, mixing TIFF and netcdf in one product is not recommended
     path: some.nc
     layer: some_var  # str: netcdf variable to read
```

UUID (Ідентифікатор універсальних унікальних ідентифікаторів) це 128-бітні значення, які використовуються в базах даних, серед іншого, для унікальної ідентифікації записів таблиці. Кожен датасет має мати свій унікальний id

Джерело: https://opendatacube.readthedocs.io/en/latest/installation/dataset-documents.html

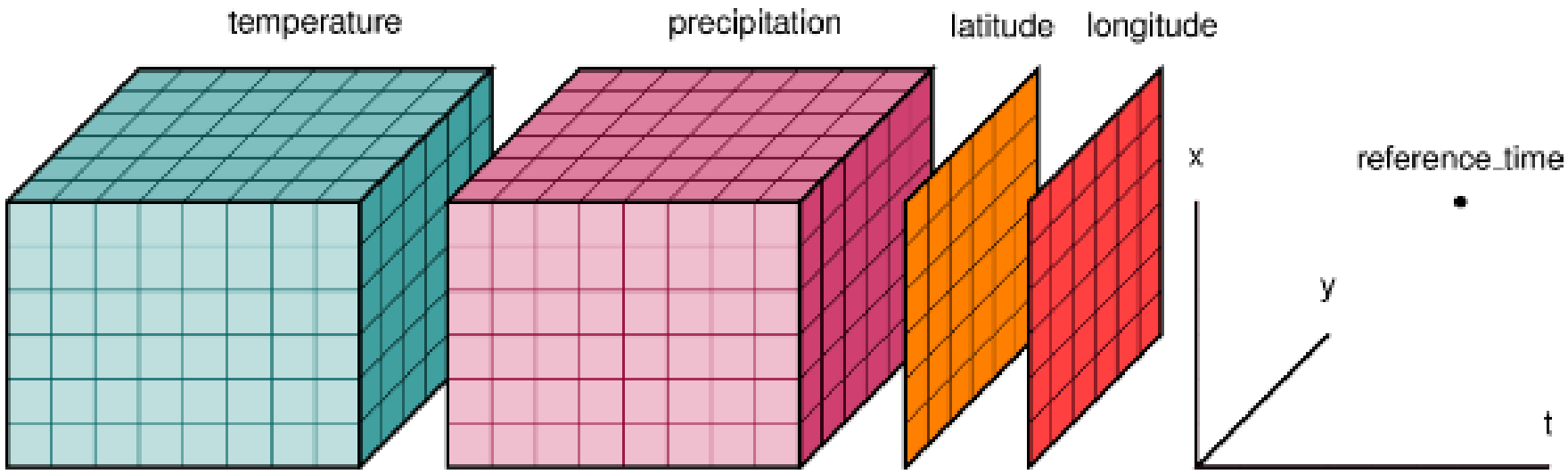
Формати геопросторових даних, сумістні з ODC

COG(Cloud Optimized GeoTIFF) - це спеціалізований формат для зберігання геопросторових зображень. Це розширення традиційного формату TIFF, оптимізоване для зберігання та отримання даних у хмарі. COG дозволяє ефективно читати частини зображень через Інтернет, що означає, що користувачі можуть отримати доступ лише до тих частин зображення, які їм потрібні, не завантажуючи весь файл. Це робить COG ідеальним форматом для роботи з великими супутниковими зображеннями в хмарних додатках.

NetCDF4 (Network Common Data Form) - це набір програмних бібліотек і самоописувальних, незалежних від машин форматів даних, які підтримують створення, доступ і обмін науковими даними, орієнтованими на масиви. NetCDF4 — це остання версія, яка вводить кілька передових особливостей, таких як більш потужні типи даних, стиснення даних та покращену продуктивність вводу/виводу. Широко використовується у науковій спільноті, особливо у кліматології, метеорології та океанографії.

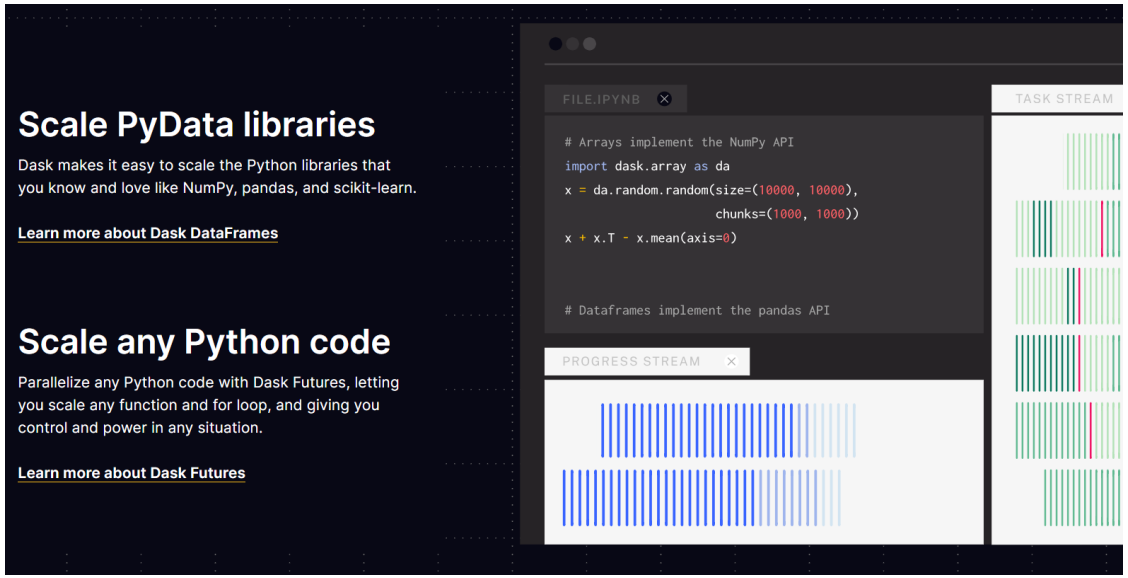
Растрові геодані можуть зберігатися як локально, так і на хмарних сховищах. Спосіб зберігання даних і їх формат потрібно враховувати під час індексації, оскільки від цього залежить формат запису шляху у файлах із метаданими.

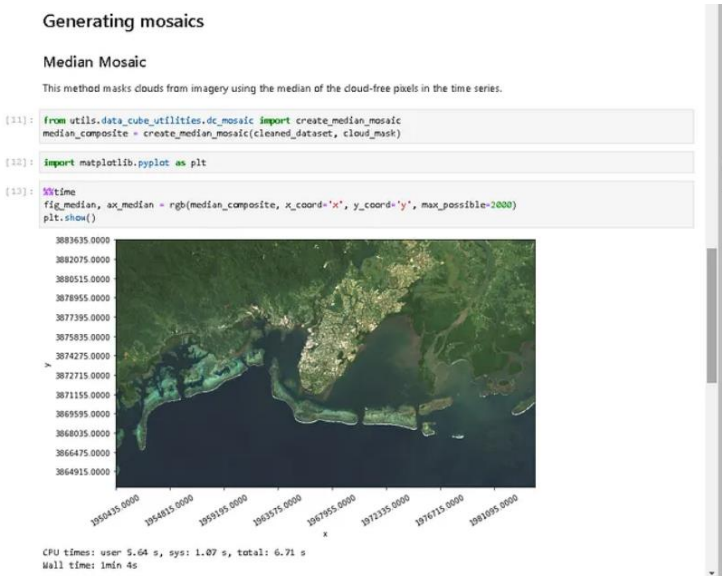
Знайомство з роботою з Big Geodata за допомогою Xarray

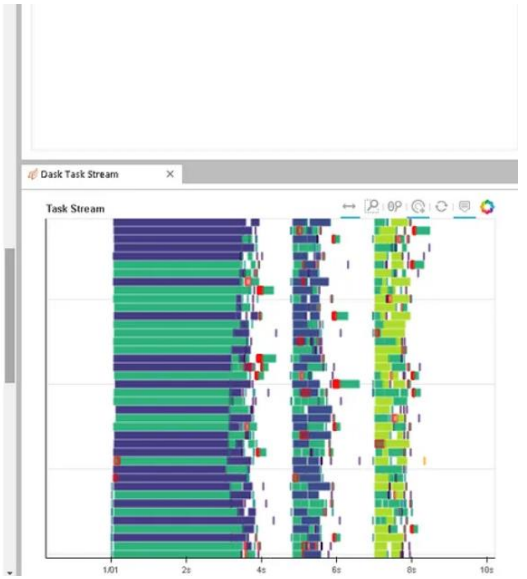


Джерело: https://docs.dea.ga.gov.au/notebooks/Beginners_guide/08_Intro_to_xarray

Знайомство із використанням Dask для оптимізації обчислювальних процесів







Джерела:

1 - https://www.dask.org;

2 - https://medium.com/@luigidifraia/adding-support-for-medians-over-dask-arrays-to-the-data-cube-utilities-7cb7faeae2

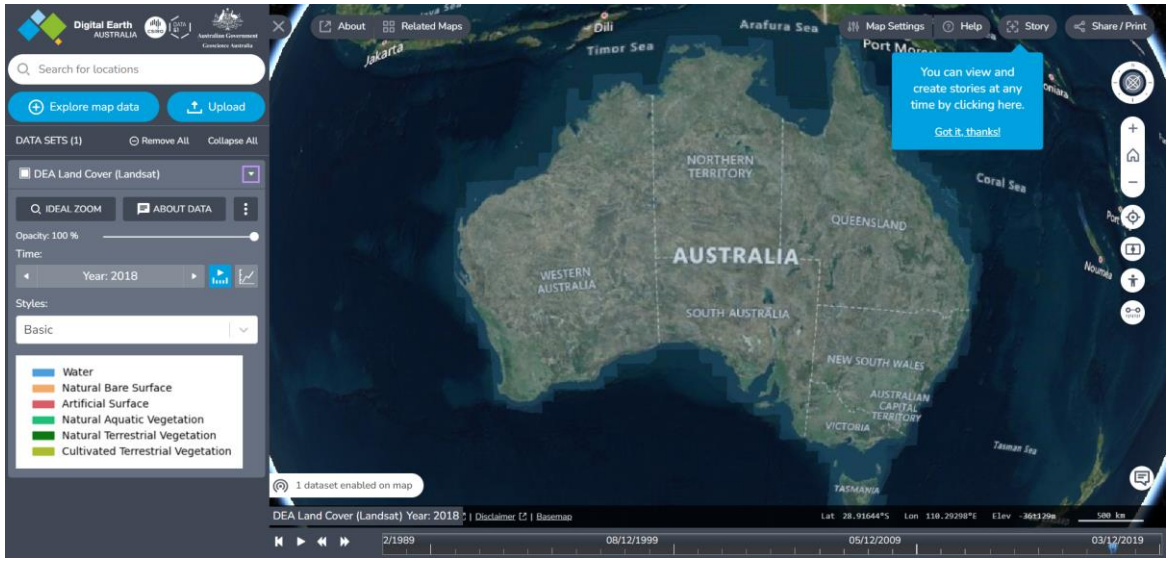
3 - https://datacube-core.readthedocs.io/en/latest/api/utilities/dask.html#

Знайомство із хмарним сервісом AWS - EC2 та S3

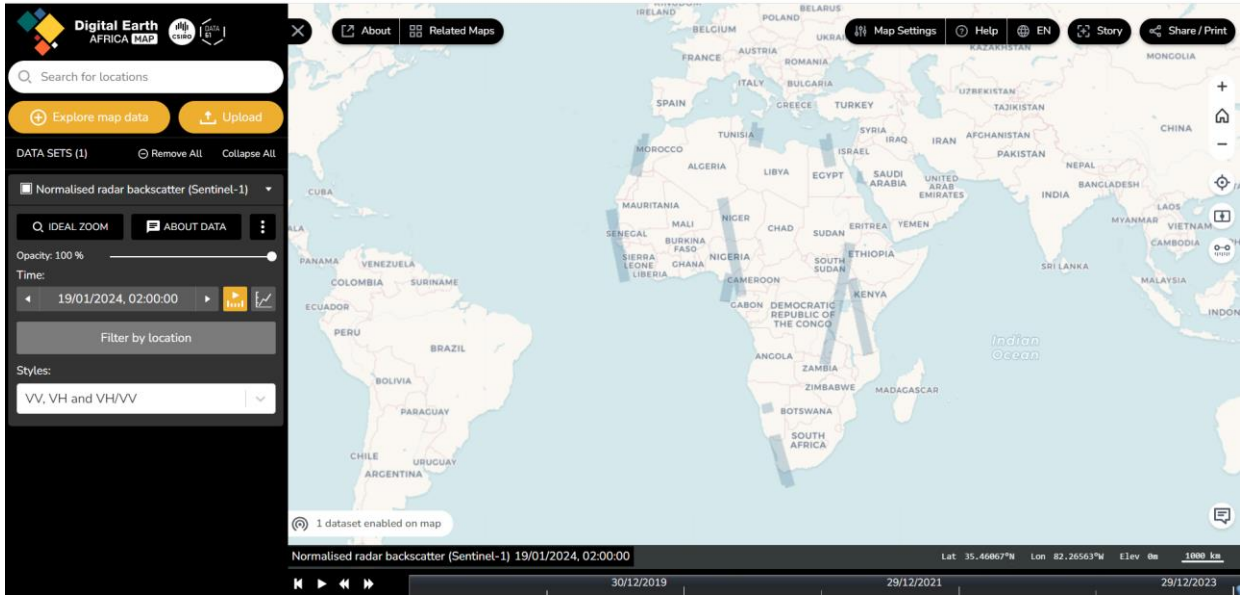
EC2 (Elastic Compute Cloud) є ключовим компонентом Amazon Web Services (AWS) і представляє собою сервіс, що надає масштабовані обчислювальні можливості в хмарі
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html

Amazon S3 (Simple Storage Service) є об'єктним сховищем, що надається Amazon Web Services (AWS) і призначений для забезпечення масштабованості, безпеки та високої доступності даних
https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html

Приклади веб-порталів реалізовані на базі Django + ODC



https://maps.dea.ga.gov.au/story/DEALandCover



https://maps.digitalearth.africa

У разі виникнення складнощів із реалізацією даної технології, питання можна адресувати ODC спільноті: http://slack.opendatacube.org

Якщо у вас є побажання, щодо покращення, їх можна надсилати у вигляді pull request на GitHub репозиторій - https://github.com/opendatacube