

GISC Tokyo cloud project / Cloud best practices

Tatsuya Noyori

Information and Communications Technology Division
Information Infrastructure Department
Japan Meteorological Agency



Cloud

[CGMS/WMO Cloud Workshop](#)

- The following inequalities hold for cost, maintenance cost and security risk.

serverless computing on cloud < virtual machine on cloud < machine on premise

- The cost based on only the time and memory allocated to execute **application code**
- **does not need** security patching work
- The cost based on the time and memory allocated to execute **virtual machine and application**
- **This means that idle time will also be charged**
- **needs** security patching work



Open Data on Cloud storage

Registry of Open Data on AWS

- Satellite
 - JMA Himawari
: <https://registry.opendata.aws/noaa-himawari/>
 - NOAA GOES
: <https://registry.opendata.aws/noaa-goes/>
 - NOAA SST: <https://registry.opendata.aws/mur/>
- Numerical Forecast Model
 - NOAA : <https://registry.opendata.aws/noaa-gfs-bdp-pds/>
 - UK Met Office
: <https://registry.opendata.aws/uk-met-office/>
 - Météo-France
: <https://registry.opendata.aws/meteo-france-models/>
 - FMI: <https://registry.opendata.aws/hirlam/>
- RADAR:
 - NOAA: <https://registry.opendata.aws/noaa-nexrad/>
 - FMI: <https://registry.opendata.aws/fmi-radar/>
 - Colombia: <https://registry.opendata.aws/ideam-radares/>
- Reanalysis
 - ECMWF : <https://registry.opendata.aws/ecmwf-era5/>
- Weather observation data
 - NOAA: <https://registry.opendata.aws/noaa-isd/>
 - CWB: https://registry.opendata.aws/cwb_opendata/



JMA Himawari

- Resource type: **S3 Bucket**
- AWS Regin: **us-east-1**
- Explore: [Browse Bucket](#)

- Resource type: **SNS Topic**
- AWS Regin: **us-east-1**
- Description: New data notifications
, only Lambda and SQS

Resources on AWS

Description

Himawari-8 Imagery

Resource type

S3 Bucket

Amazon Resource Name (ARN)

```
arn:aws:s3:::noaa-himawari8
```

AWS Region

```
us-east-1
```

AWS CLI Access (No AWS account required)

```
aws s3 ls s3://noaa-himawari8/ --no-sign-request
```

Explore

[Browse Bucket](#)

Description

New data notifications for Himawari-8, only Lambda and SQS protocols allowed

Resource type

SNS Topic

Amazon Resource Name (ARN)

```
arn:aws:sns:us-east-1:123901341784:NewHimawari8Object
```

AWS Region

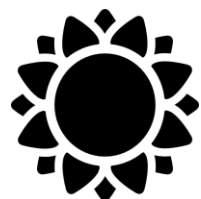
```
us-east-1
```



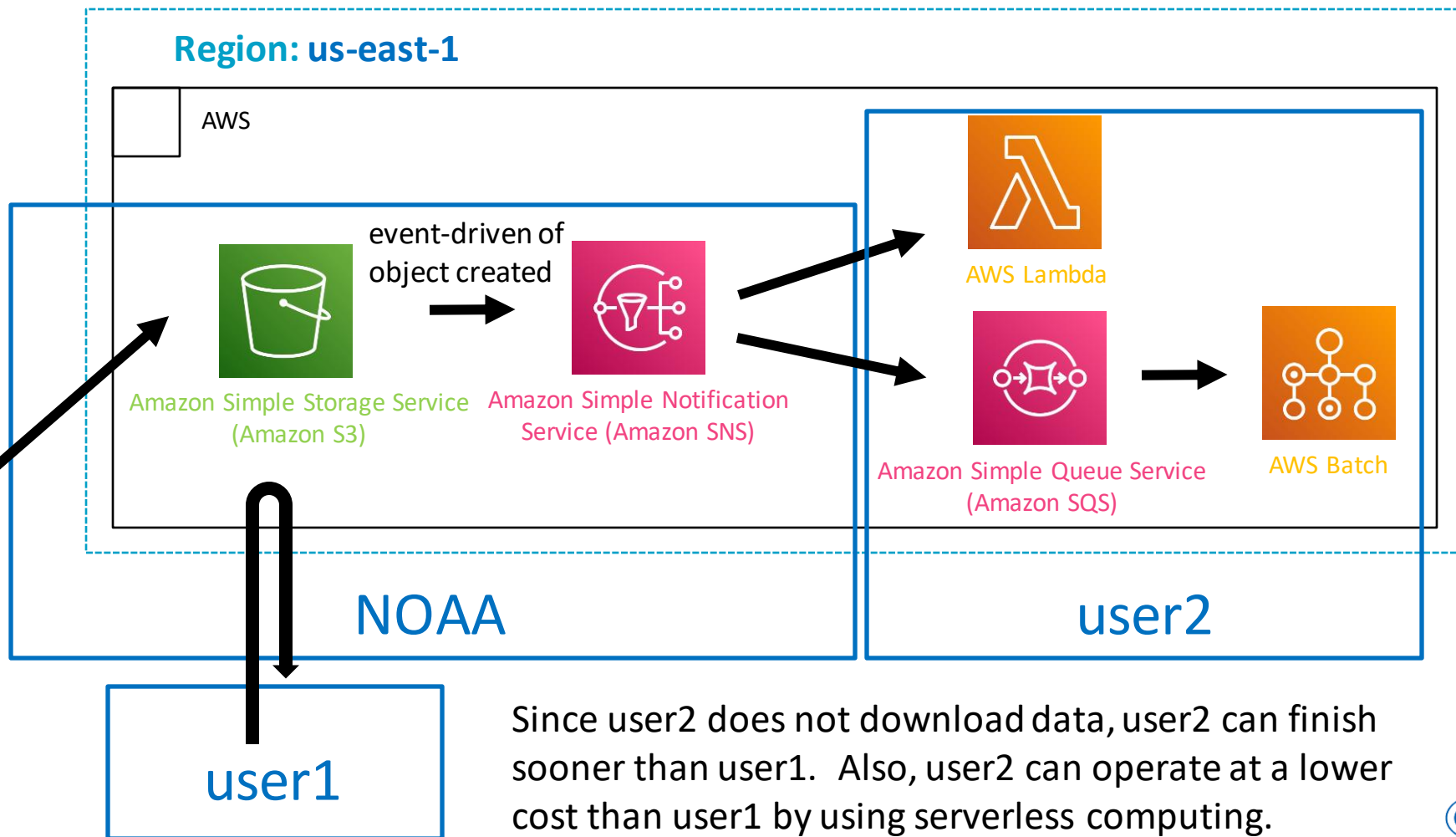


JMA Himawari

JMA Himawari



NOAA/
NESDIS



Since user2 does not download data, user2 can finish sooner than user1. Also, user2 can operate at a lower cost than user1 by using serverless computing.





Open Data on AWS / Open Data Sponsorship Program

The Amazon Web Services (AWS) Open Data Sponsorship Program covers the cost of storage for publicly available high-value cloud-optimized datasets. We work with data providers who seek to:

- Democratize access to data by making it available for analysis on AWS
- Develop new cloud-native techniques, formats, and tools that lower the cost of working with data
- Encourage the development of communities that benefit from access to shared datasets

The following data are only S3 Bucket resource without Amazon SNS resource.

- Colombia: <https://registry.opendata.aws/ideam-radares/>
- ECMWF : <https://registry.opendata.aws/ecmwf-era5/>
- CWB: https://registry.opendata.aws/cwb_opendata/

The cost would be lower if used only S3 bucket with Open Data Sponsorship Program.



data copy using index files on cloud storage only (1 message = 1 file)

- Synchronization between cloud storage

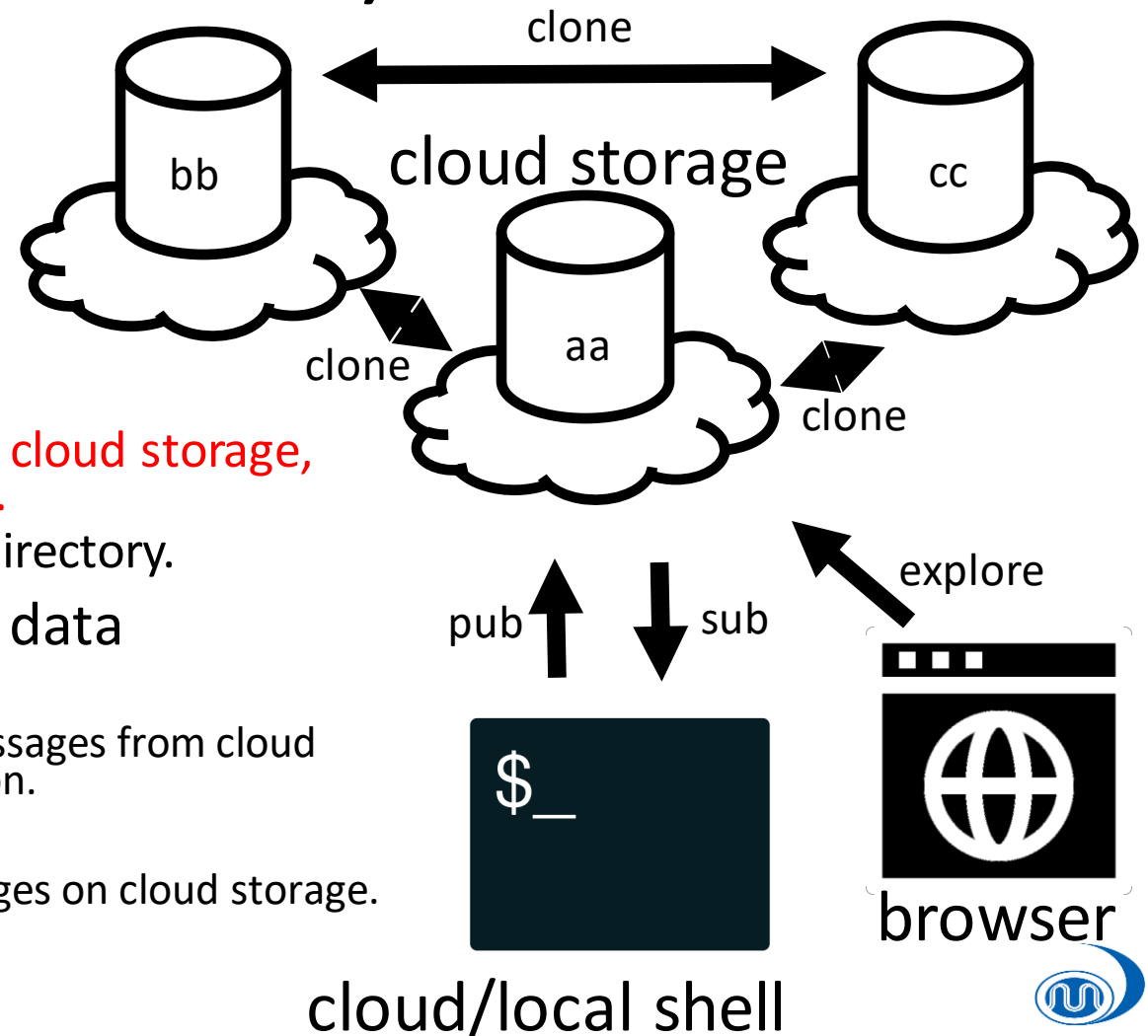
- [aa cloud storage explore](#)
- [bb cloud storage explore](#)
- [cc cloud storage explore](#)

By placing index files describing the created file in the cloud storage, users can synchronize, subscribe, and search the data.

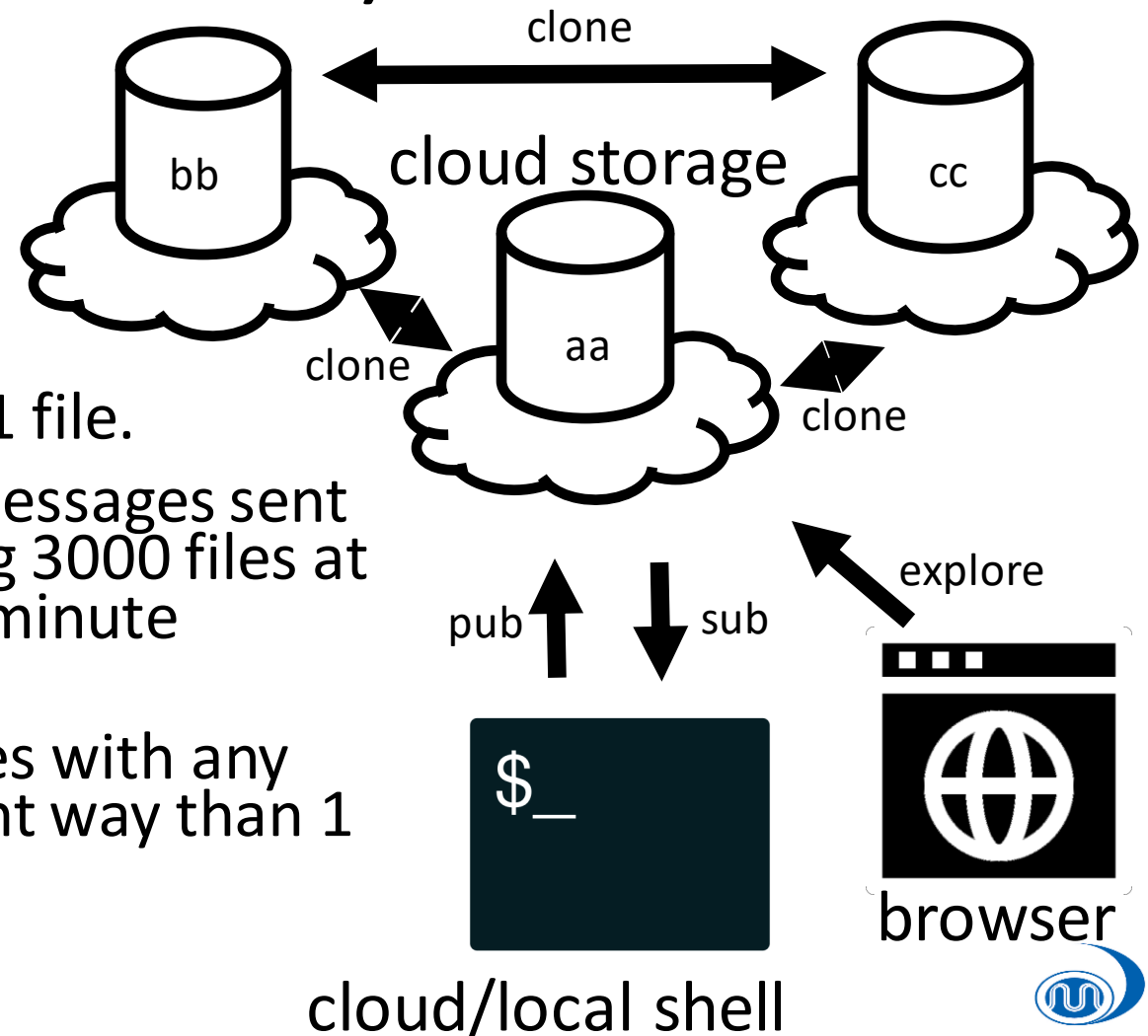
The index files are placed in /4PubSub and /4Search directory.

- publish/subscribe/clone and search/download data

- [How to subscribe/search/download](#)
 - If you follow the procedure, you can download GTS messages from cloud storage. This is like the current WIS subscription function.
- [How to publish/clone](#)
 - If you follow the procedure, you can upload GTS messages on cloud storage.



many small file copy is inefficient (1 message = 1 file)



- The above implementation is 1 message = 1 file.
- NWP/Satellite and observation are many messages sent at once. EUSR/bufr/satellite/siral is creating 3000 files at once with a small size of 400bytes. Also, 1 minute data is divided into 60 files.
- Since it is inefficient to copy many small files with any protocols, it is better to find a more efficient way than 1 message = 1 file.

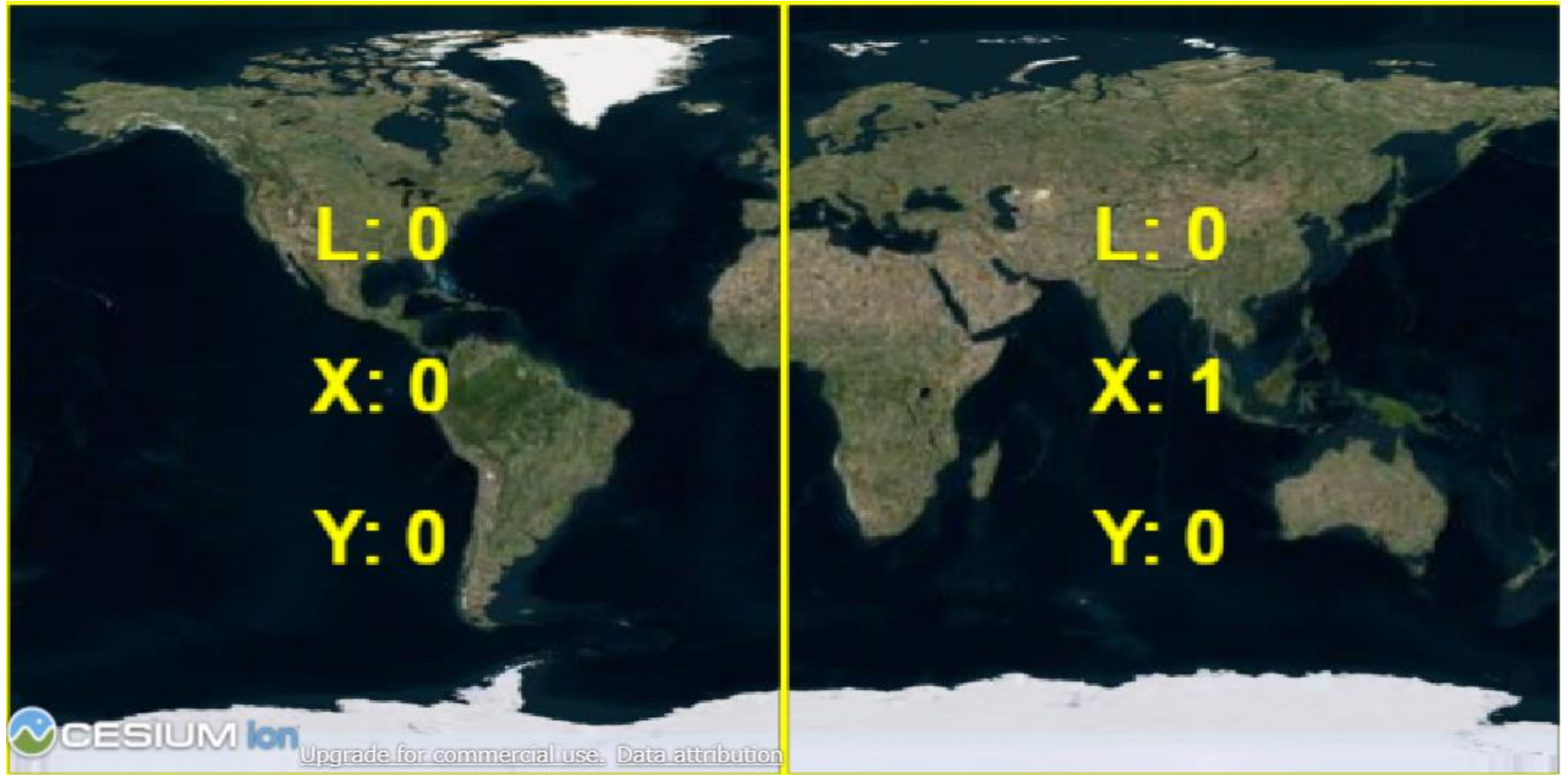


Tile dataset

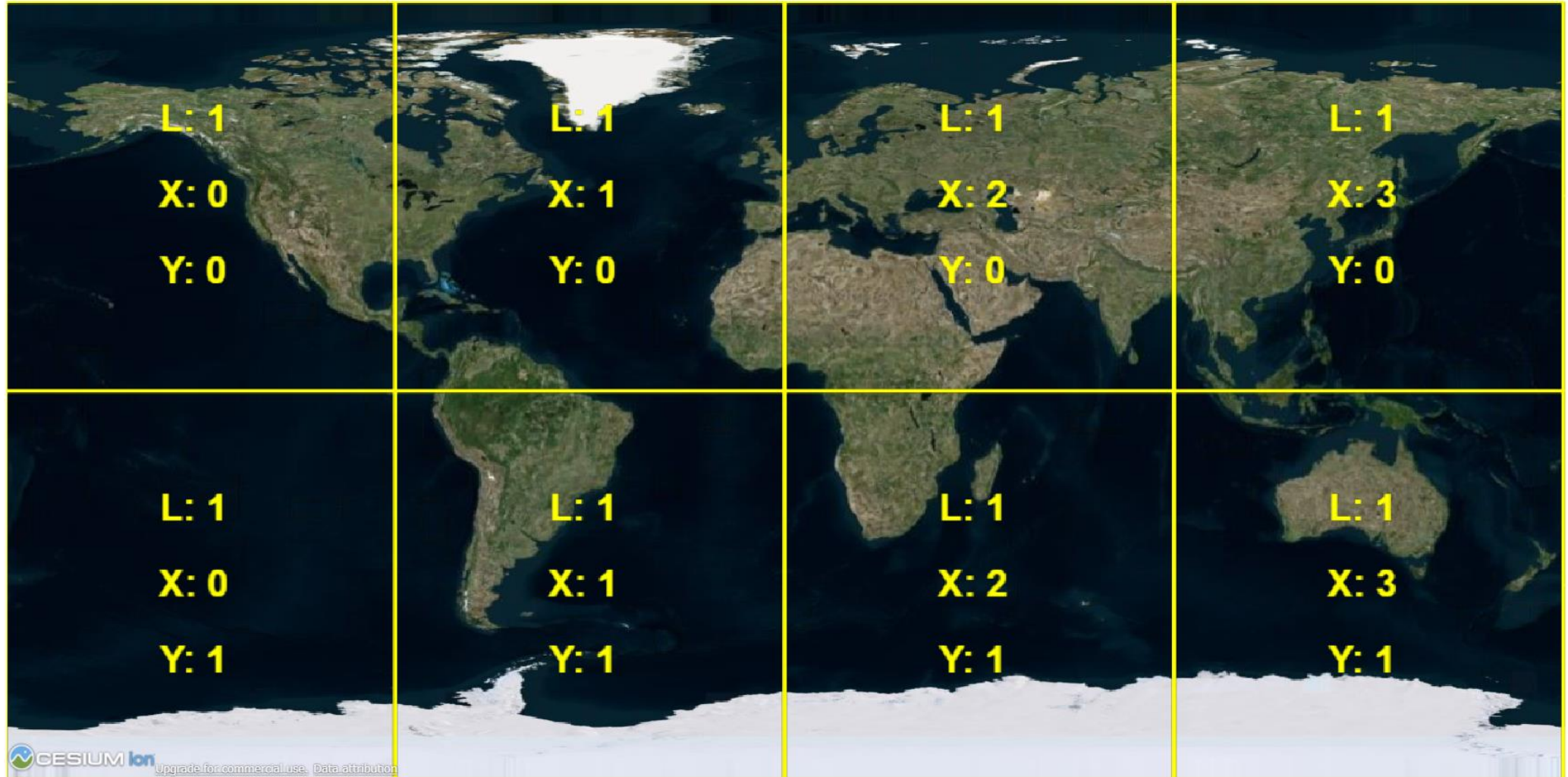
- dataset?
 - [a collection of data](#)
 - many records in 1 file (Not 1 message = 1 file) (many = 1 ~ 1000000)
 - 1 file size = KB ~ MB
- The following is tile dataset.
<http://202.32.195.138:9000/aa-open-dataset/4Site/explore.html>
- tile?
 - [tile map](#)
 - A Tile is a data file in the area where the earth is divided into tiles.



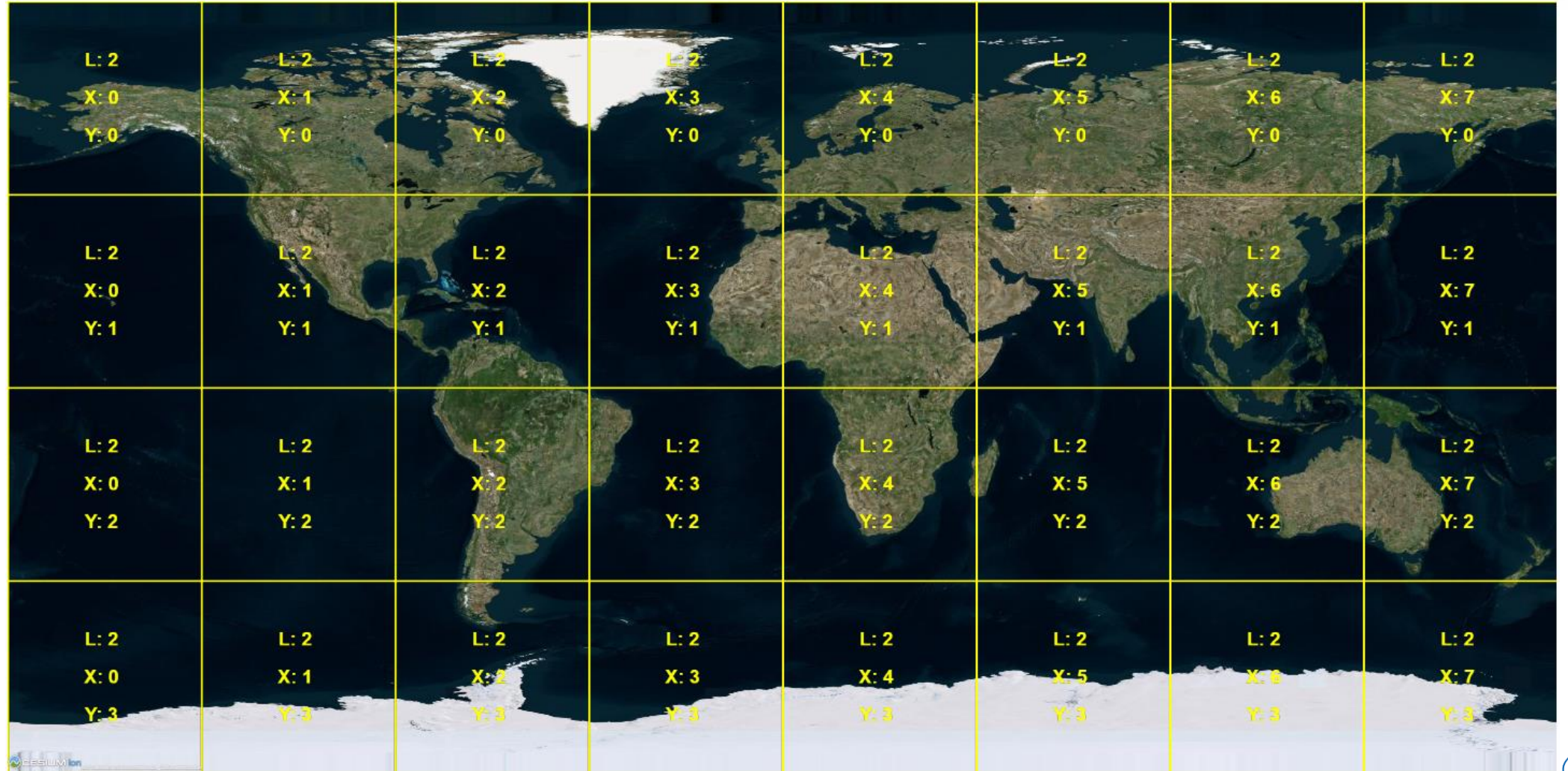
Tile Level 0 = 2 tiles



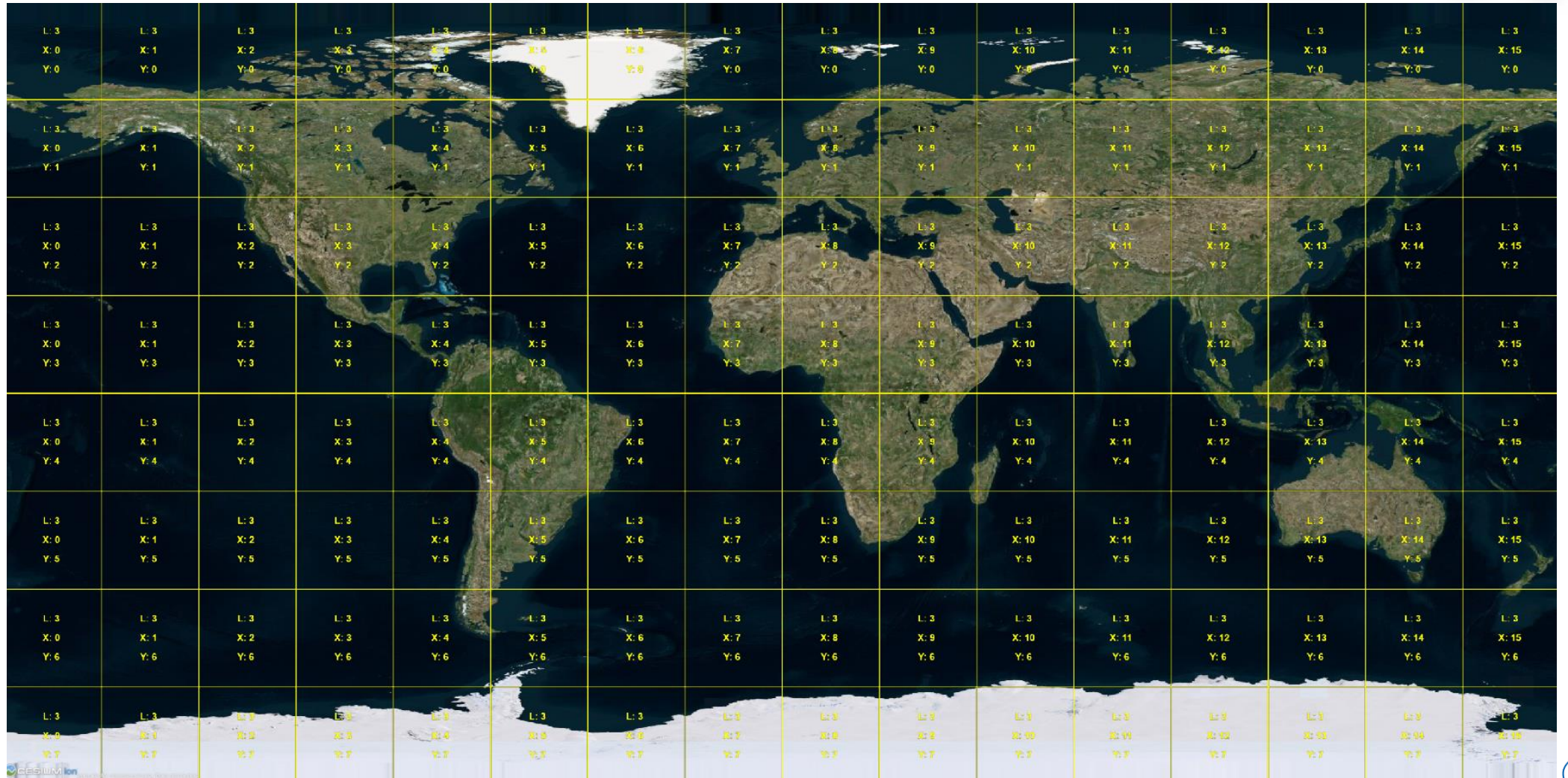
Tile Level 1 = 8 tiles (Ocean/Surface/Upper_air)



Tile Level 2 = 32 tiles



Tile Level 3 = 128 tiles (Satellite/NWP)



tile dataset

- [Explore of tile dataset](#)
 - directory path for all : aa-open-dataset/4All/bufr_to_arrow/ocean/float/2022/0111/0030/l1x1y0.arrow
 - directory path for CCCC: aa-open-dataset/CCCC/bufr_to_arrow/ocean/float/2022/0110/0630/l1x3y0.arrow
- One tile file of the tile dataset contains 10 minutes data.
- Directory path: category/subcategory/year/day+month/hour+minute/
- Category/Subcategory: observation space/observation method
- File name: l\${tile_level}x\${tile_x}y\${tile_y}.arrow
- Update file: Tile files are updated with the latest merged data.
- File format: [Apache Arrow](#)
- File compression: gzip/brotli for HTTP/HTTPS Encoding
- Subscribe/Search: users can subscribe/search by using index files
 - If you get the data when you need it, you can get the latest data at that time.
- Select/get tile: tile can be selected by URL path.



tile dataset viewer

- tile dataset viewer

http://202.32.195.138:9000/aa-open-dataset/4Site/tile_dataset_viewer.html

- tile dataset viewer for Asia

http://202.32.195.138:9000/aa-open-dataset/4Site/tile_dataset_viewer_for_asia.html

is filtered with Asia area tiles 'l1x2y0', 'l1x2y1', 'l1x3y0', 'l1x3y1',

'l2x5y0', 'l2x6y0', 'l2x7y0', 'l2x5y1', 'l2x6y1', 'l2x7y1', 'l2x5y2', 'l2x6y2', 'l2x7y2',

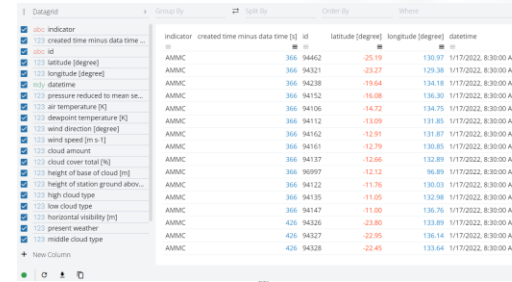
'l3x9y1', 'l3x10y1', 'l3x11y1', 'l3x12y1', 'l3x13y1', 'l3x14y1',

'l3x9y2', 'l3x10y2', 'l3x11y2', 'l3x12y2', 'l3x13y2', 'l3x14y2',

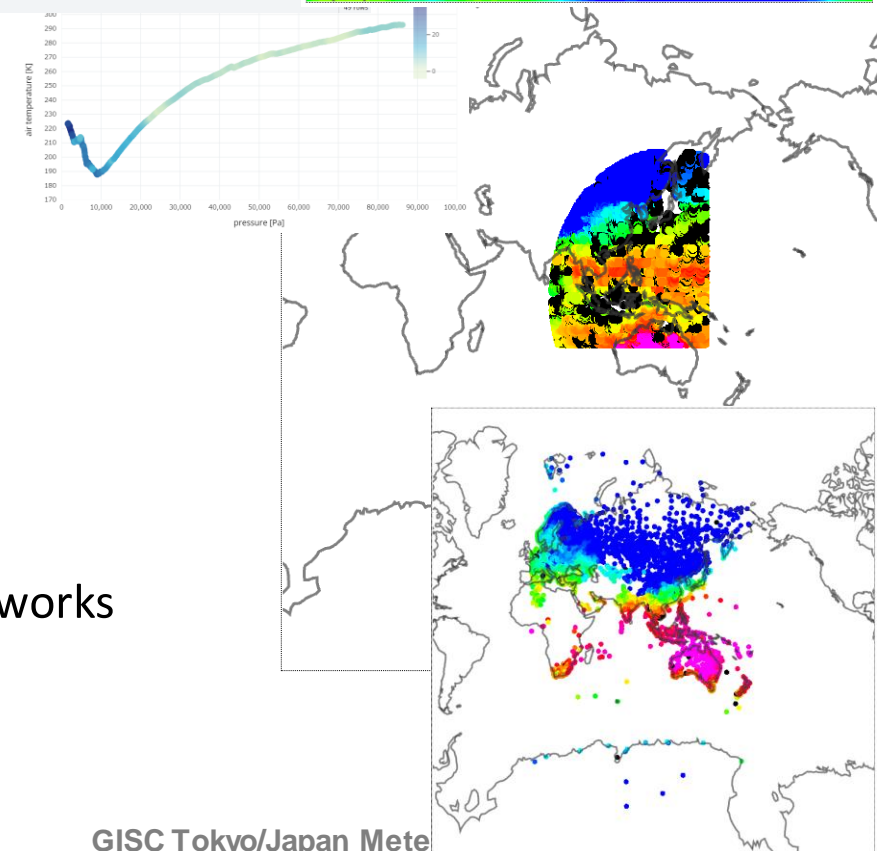
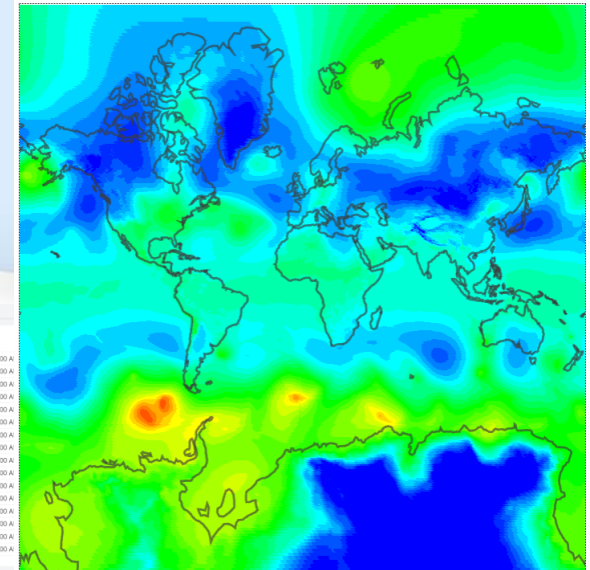
'l3x9y3', 'l3x10y3', 'l3x11y3', 'l3x12y3', 'l3x13y3', 'l3x14y3',

'l3x9y4', 'l3x10y4', 'l3x11y4', 'l3x12y4', 'l3x13y4', 'l3x14y4'

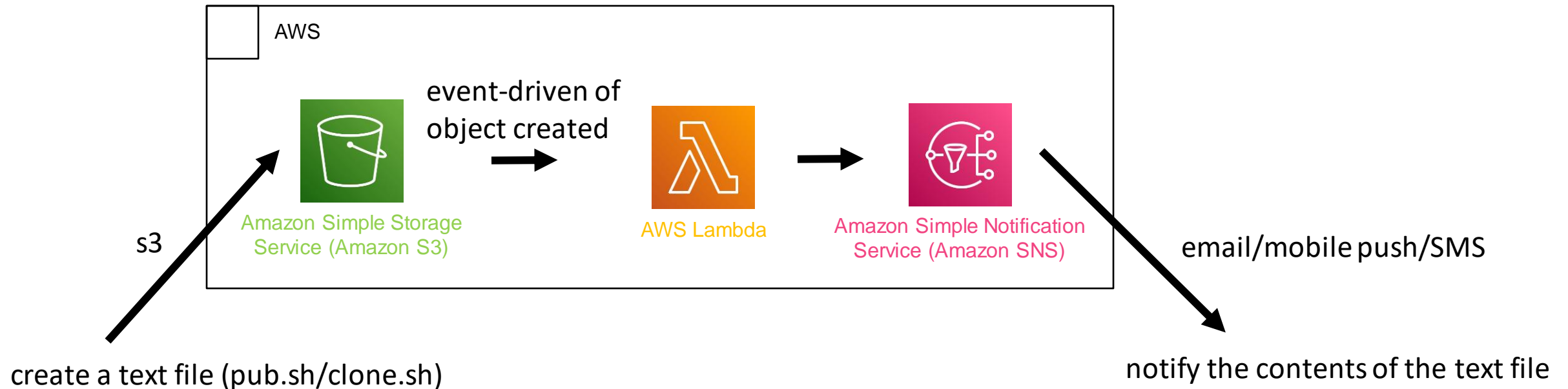
- The above application works only with javascript, so the cost is low since it works only with S3 Bucket.



Indicator	created time minus data time [s]	lat	longitude [degree]	datetime
AMMC	366 94462	-25.19	130.97	1/17/2022, 8:30:00 A
AMMC	366 94321	-23.27	129.38	1/17/2022, 8:30:00 A
AMMC	366 94238	-19.64	130.16	1/17/2022, 8:30:00 A
AMMC	366 94152	-16.08	136.30	1/17/2022, 8:30:00 A
AMMC	366 94106	-14.72	134.75	1/17/2022, 8:30:00 A
AMMC	366 94112	-13.09	131.85	1/17/2022, 8:30:00 A
AMMC	366 94162	-12.91	131.87	1/17/2022, 8:30:00 A
AMMC	366 94161	-12.79	130.85	1/17/2022, 8:30:00 A
AMMC	366 94137	-12.66	132.89	1/17/2022, 8:30:00 A
AMMC	366 96997	-12.12	96.89	1/17/2022, 8:30:00 A
AMMC	366 94122	-11.76	130.03	1/17/2022, 8:30:00 A
AMMC	366 94195	-11.05	132.98	1/17/2022, 8:30:00 A
AMMC	366 94147	-11.00	130.76	1/17/2022, 8:30:00 A
AMMC	426 94326	-23.80	133.89	1/17/2022, 8:30:00 A
AMMC	426 94327	-23.95	136.14	1/17/2022, 8:30:00 A
AMMC	426 94328	-22.45	133.64	1/17/2022, 8:30:00 A



An example of simple notification with serverless computing on cloud



A sample implementation is [here](#).



Multicloud

- Cloud does not have to be limited to one company, and the good function of each cloud can be used like utility. Select and use the good functions with low cost in each cloud.



cloud best practice

- Migrate from message exchange to dataset synchronization.
 - dataset is more efficient than transferring many small message files.
- Create index file for the created/updated files of dataset on cloud storage to make users synchronize/select easily.
- Open and Share dataset on cloud storage as open data with open data program of cloud.
- Use 10-minutes tile dataset.
 - It is more efficient to transfer 10-minutes tile dataset than to transfer many small messages divided into seconds/minutes.
 - Tile dataset makes it easy for users to select the tiles they need.
 - The tile path of category/subcategory/year/day+month/hour+minute/l?x?y?.arrow makes it easy to implement applications.



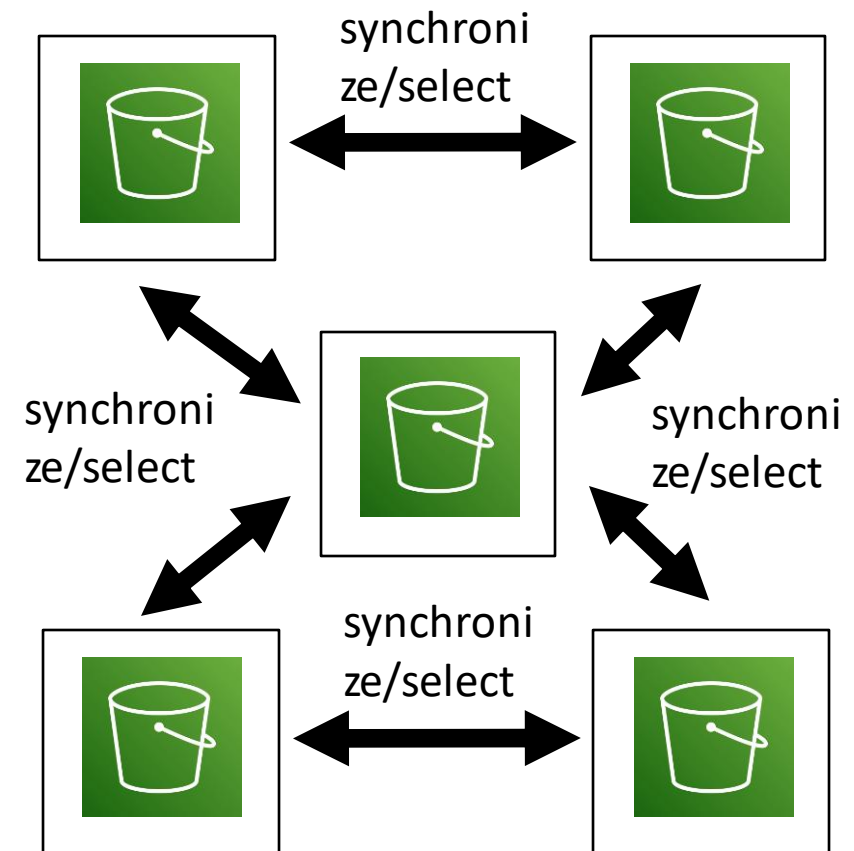
cloud best practice

- Use [Apache Arrow](#) as file format.
 - Apache Arrow is available in CPU/GPU, [AI](#) and browser.
 - Data in the Apache Arrow format can be read by Javascript in browser, so data in the Apache Arrow format can be easily shared with many users by using table view([Perspective](#)) and map view([Deck.gl-loaders.gl](#)).
 - Using the Apache arrow format, applications can be created by only Javascript.
 - Javascript-only application is more secure than server-side application.
 - Javascript-only application is lower cost because it works with only cloud storage.
- Use [Brotli](#) compression with HTTPS encoding.
- Use [AWS JavaScript S3 Explorer](#) to make users to easily browse dataset on cloud storage.
- Use notification service of cloud for emergency alerts



Independent agencies work together

- Each agency manage its own data on their cloud storage as open data.
- An agency publish/share own data on own cloud storage for own users.
- Other agencies synchronize/select the data as the user.



GitHub

- The program created in this project is available at the following link.
 - https://github.com/public-tatsuya-noyori/meteorological_preprocessor
 - https://github.com/public-tatsuya-noyori/meteorological_visualizer
- If you are interested in this development, please contact me.



Thank you

ありがとうございました。

