How to execute genome analysis on Cloud

An introduction of Extended-ETL engine: `awsub`

Hiromu OCHIAI - National Cancer Center Japan

Genome analysis on Cloud Resources

More and more people are using cloud resources to analyze their sample sequences.

and more

The best practice of "Genome Analysis on Cloud"?

1. "Building a Cluster on Cloud"

- Galaxy
- cfn-cluster
- ElastiCluster
- Butler
- etc...

1. Pros and Cons of "Cluster on Cloud"

1. Pros and Cons of "Cluster on Cloud"

- Pros:
  - We are **VERY** used to cluster on HPC
    - *Grid Engine, HTCondor, SLURM, etc...*
    - e.g. `qsub ./my-workflow.sh`

1. Pros and Cons of "Cluster on Cloud"

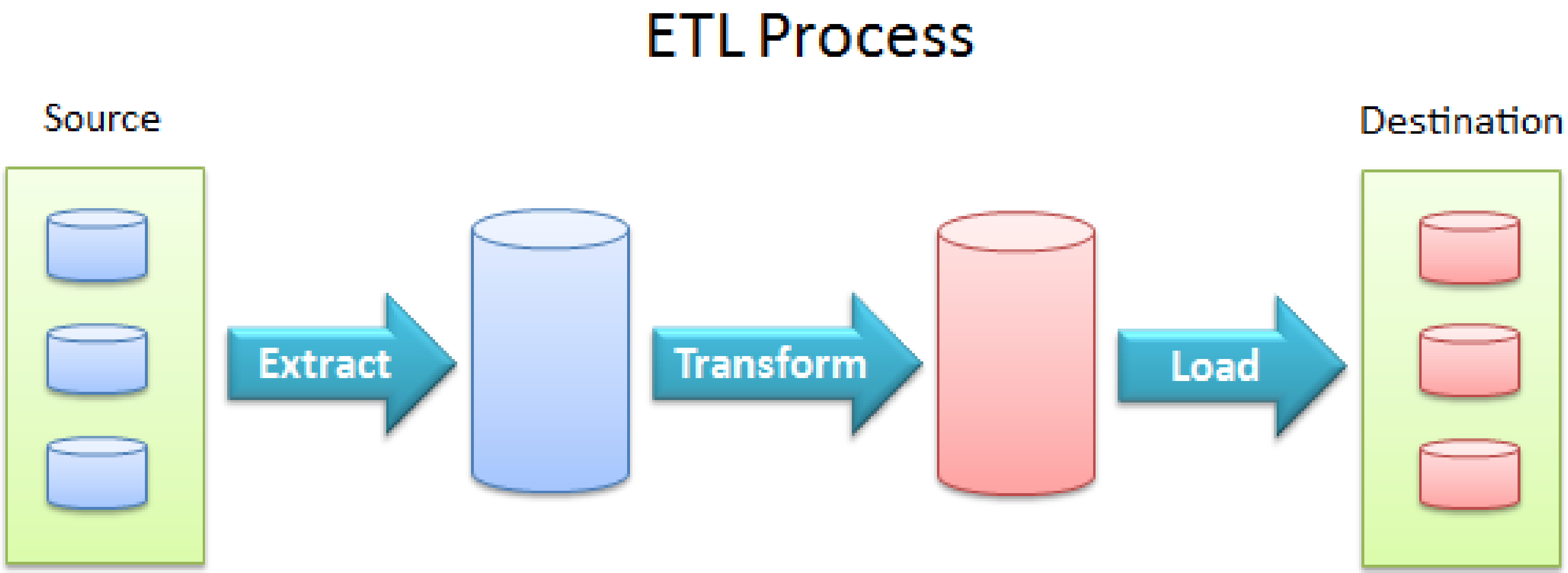
- Pros:
  - We are **VERY** used to cluster on HPC
    - *Grid Engine, HTCondor, SLURM, etc...*
    - e.g. `qsub ./my-workflow.sh`
- Cons:
  - Unnecessary instances time
  - // Inefficient shared disk I/O

2. Suggestion:

2. Suggestion: "on-demand ETL on Cloud"

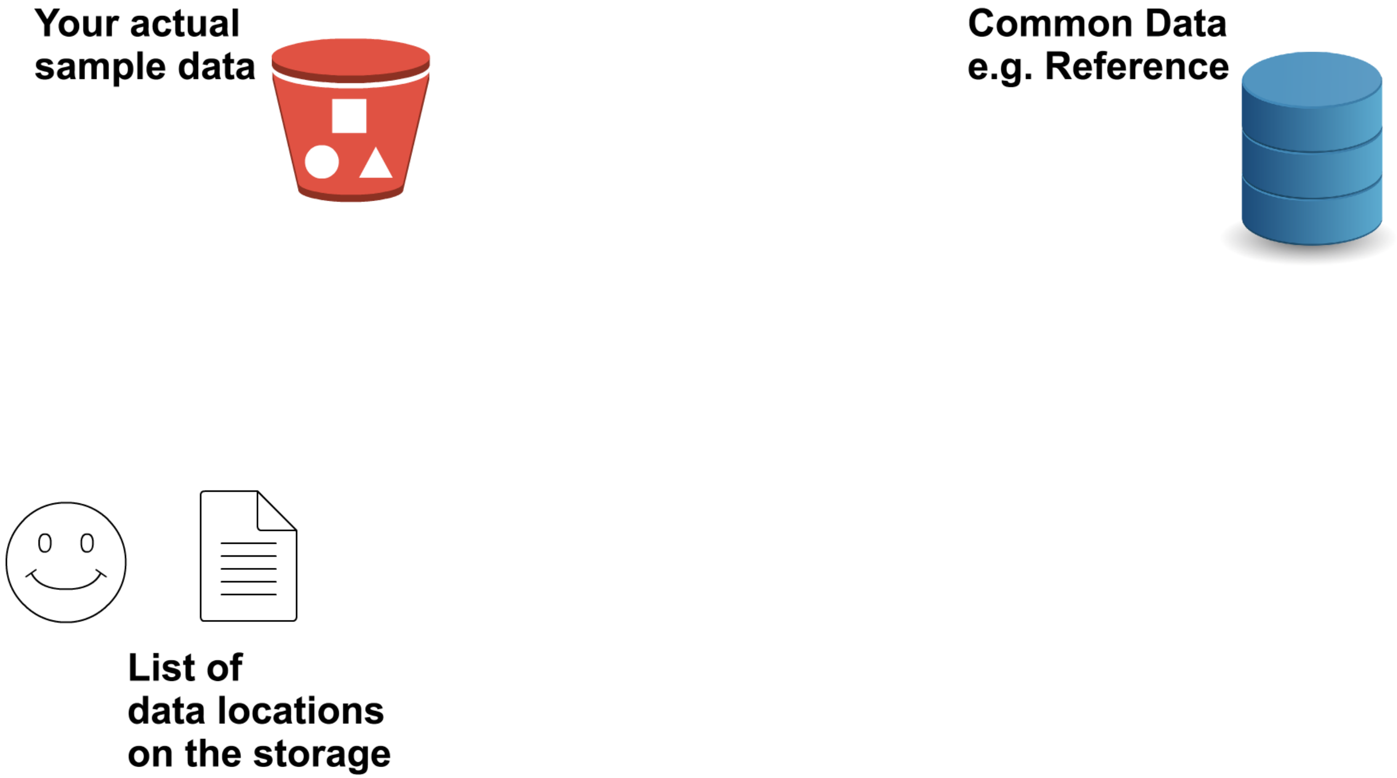
ETL is

- Extract, Transform, Load
- Data processing model for general purpose



Use Case

If you have 4 Fastq samples

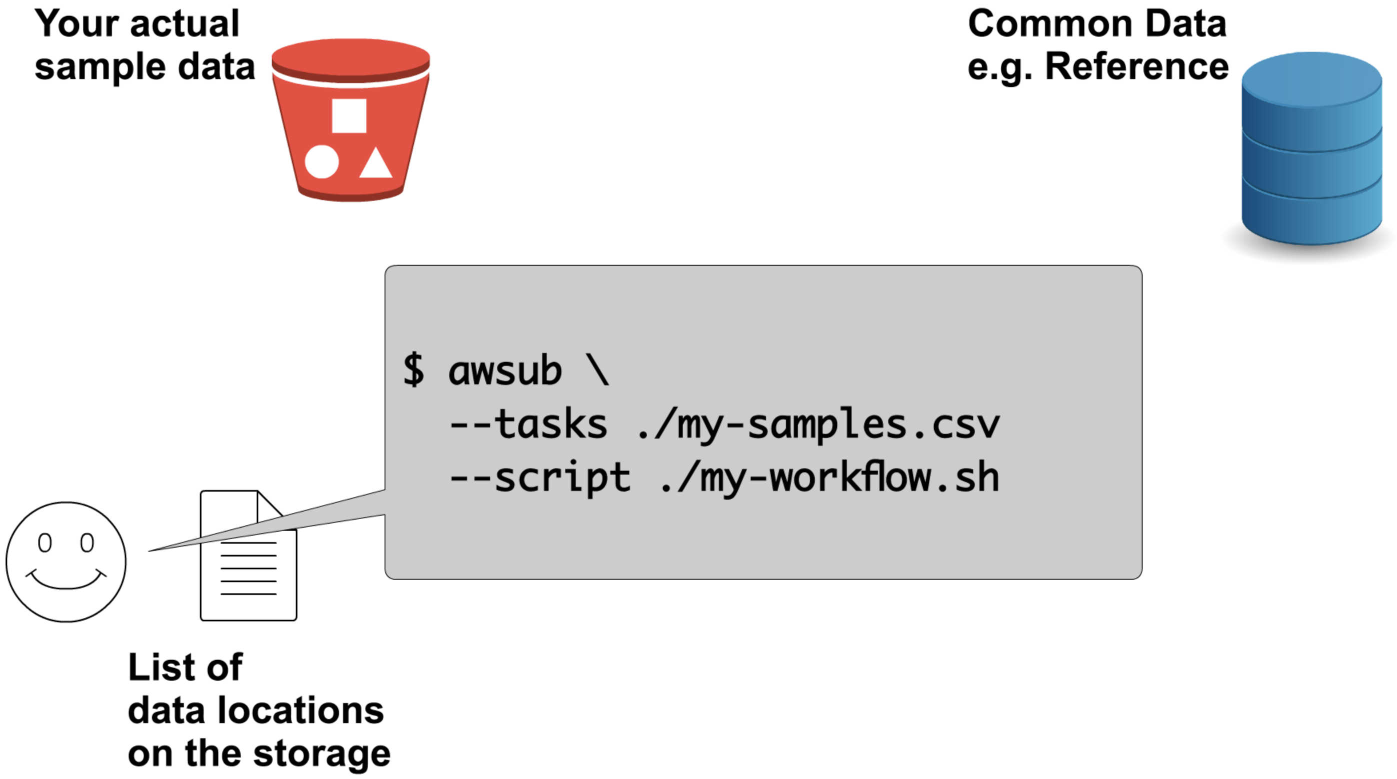


Your actual sample data

Common Data e.g. Reference

List of data locations on the storage

Specify workflow script and samples



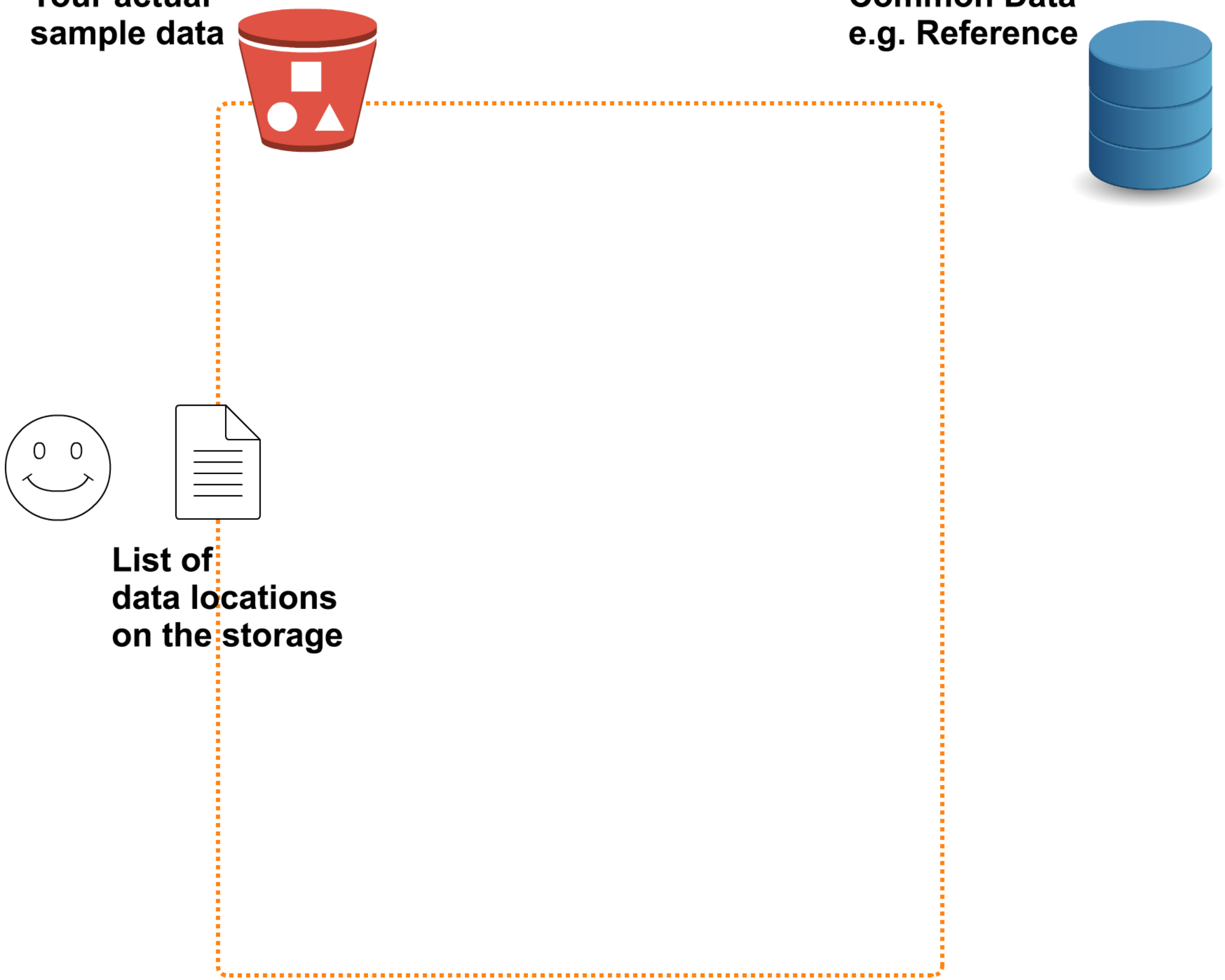
Your actual sample data

Common Data e.g. Reference

```
$ awsub \
  --tasks ./my-samples.csv
  --script ./my-workflow.sh
```

List of data locations on the storage

Security Group

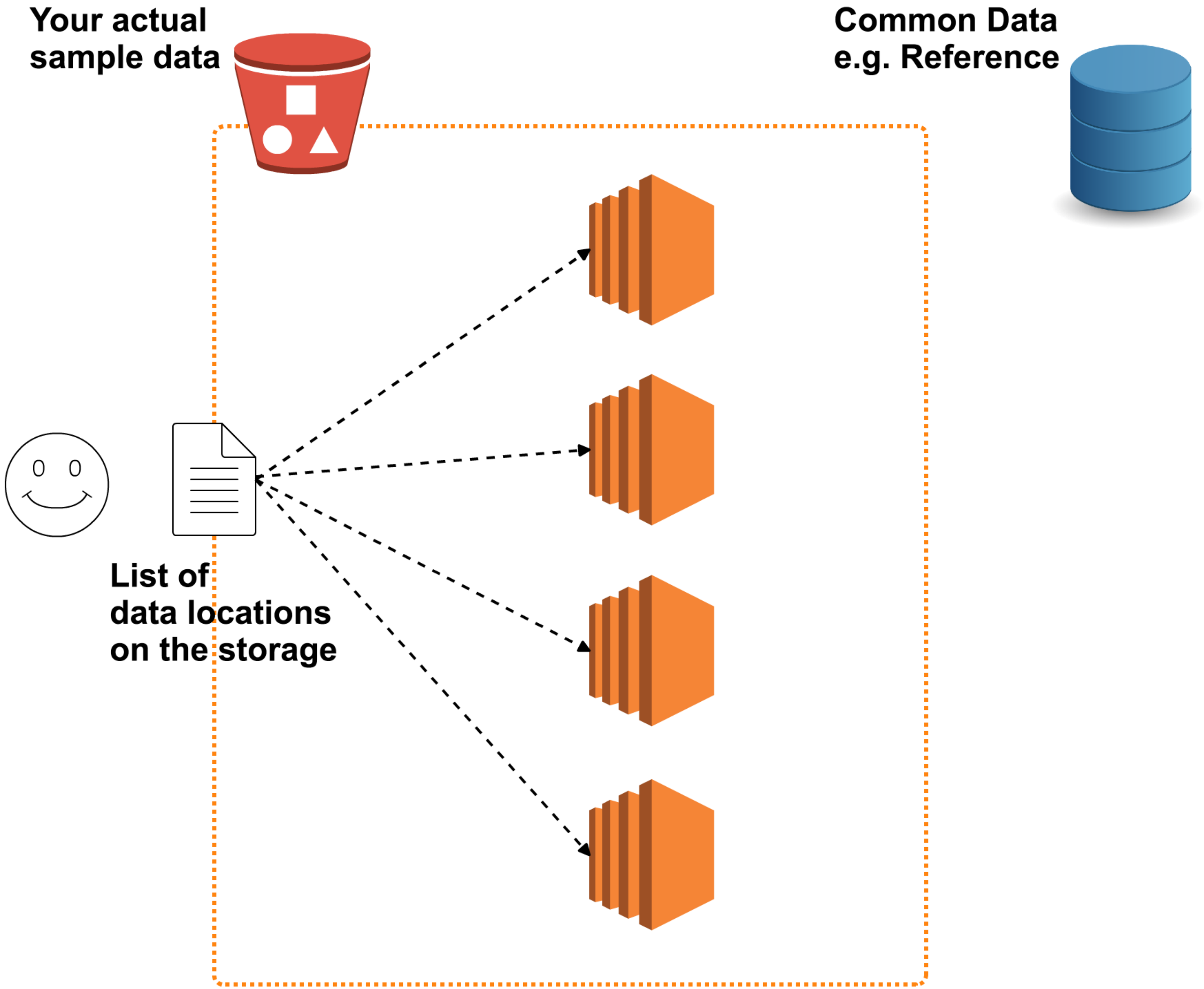


Your actual
sample data

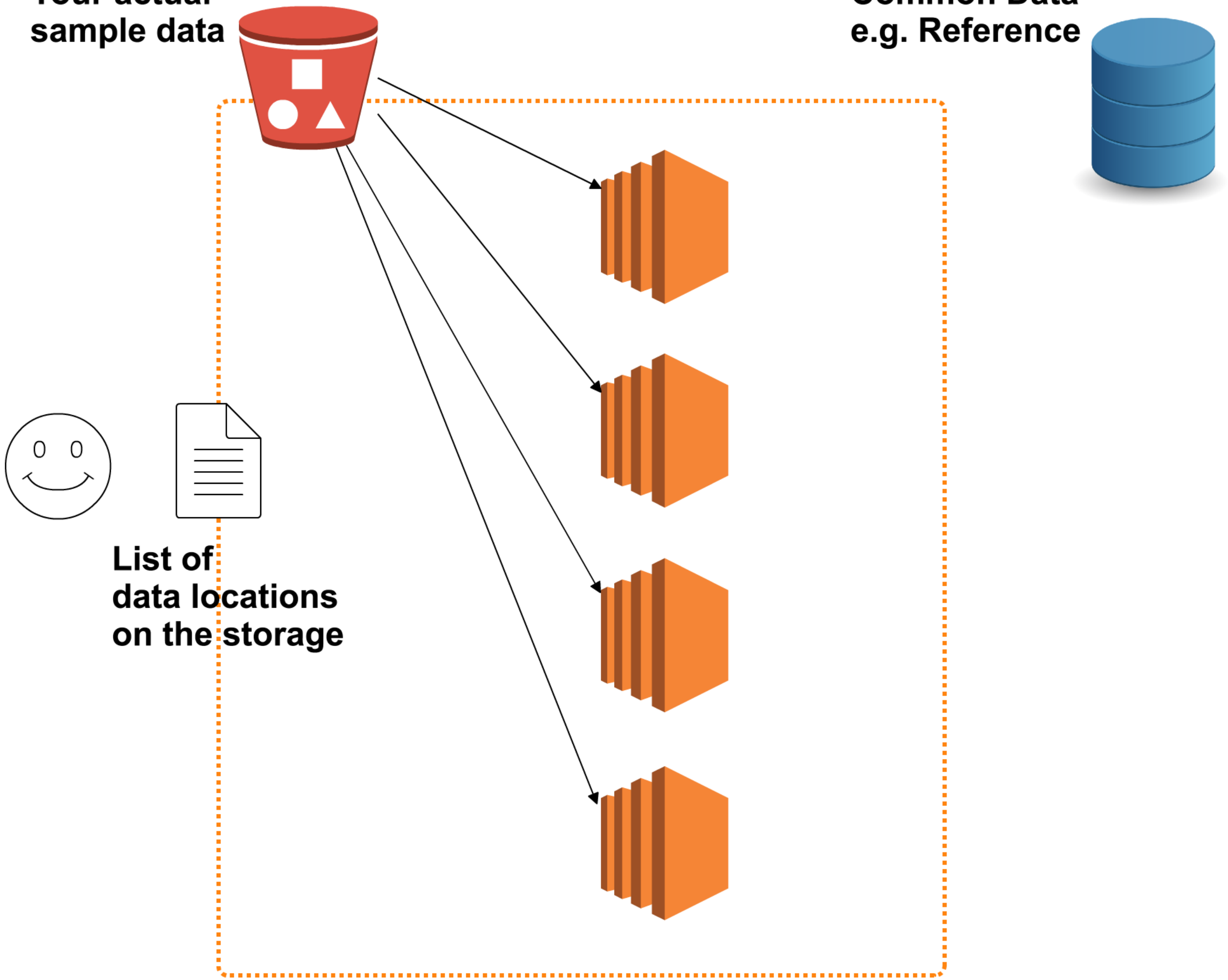
Common Data
e.g. Reference

List of
data locations
on the storage

Inscances for each sample



Fetch specific sample data according to the location



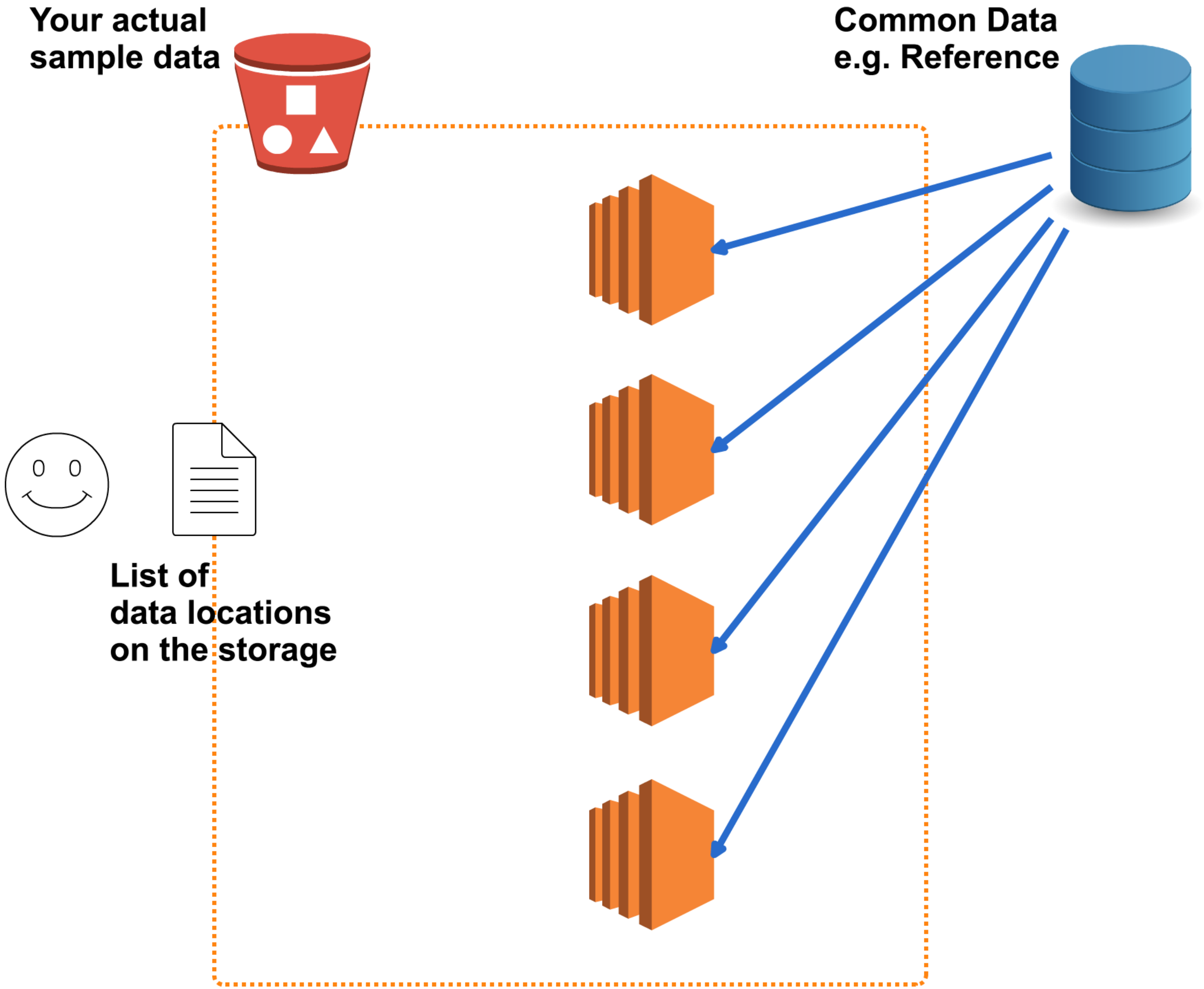
Your actual sample data

Common Data e.g. Reference

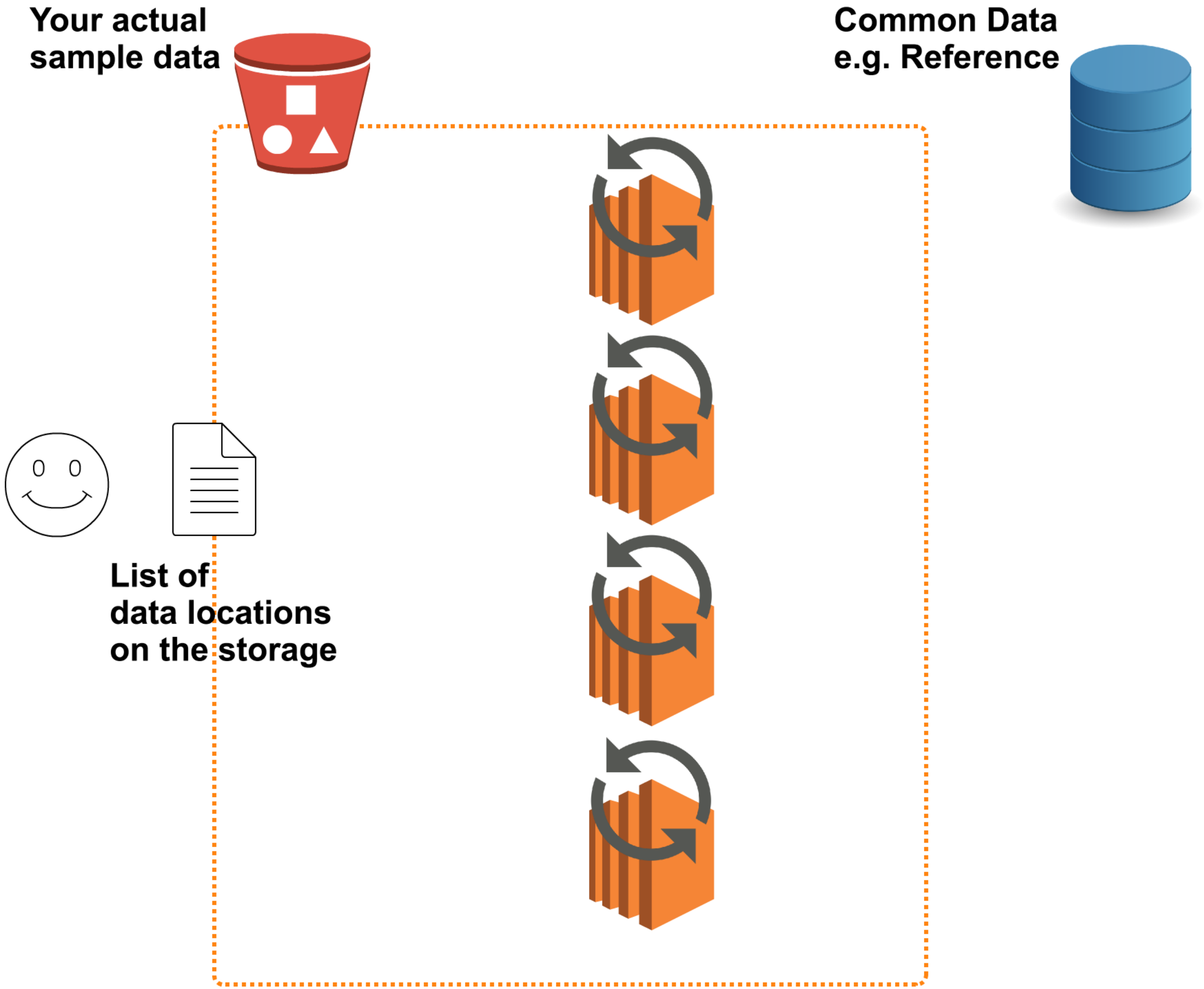
List of data locations on the storage

// それぞれ違うfastqであることをわかりやすくする

Fetch reference data from common data source



Execute your workflow for each

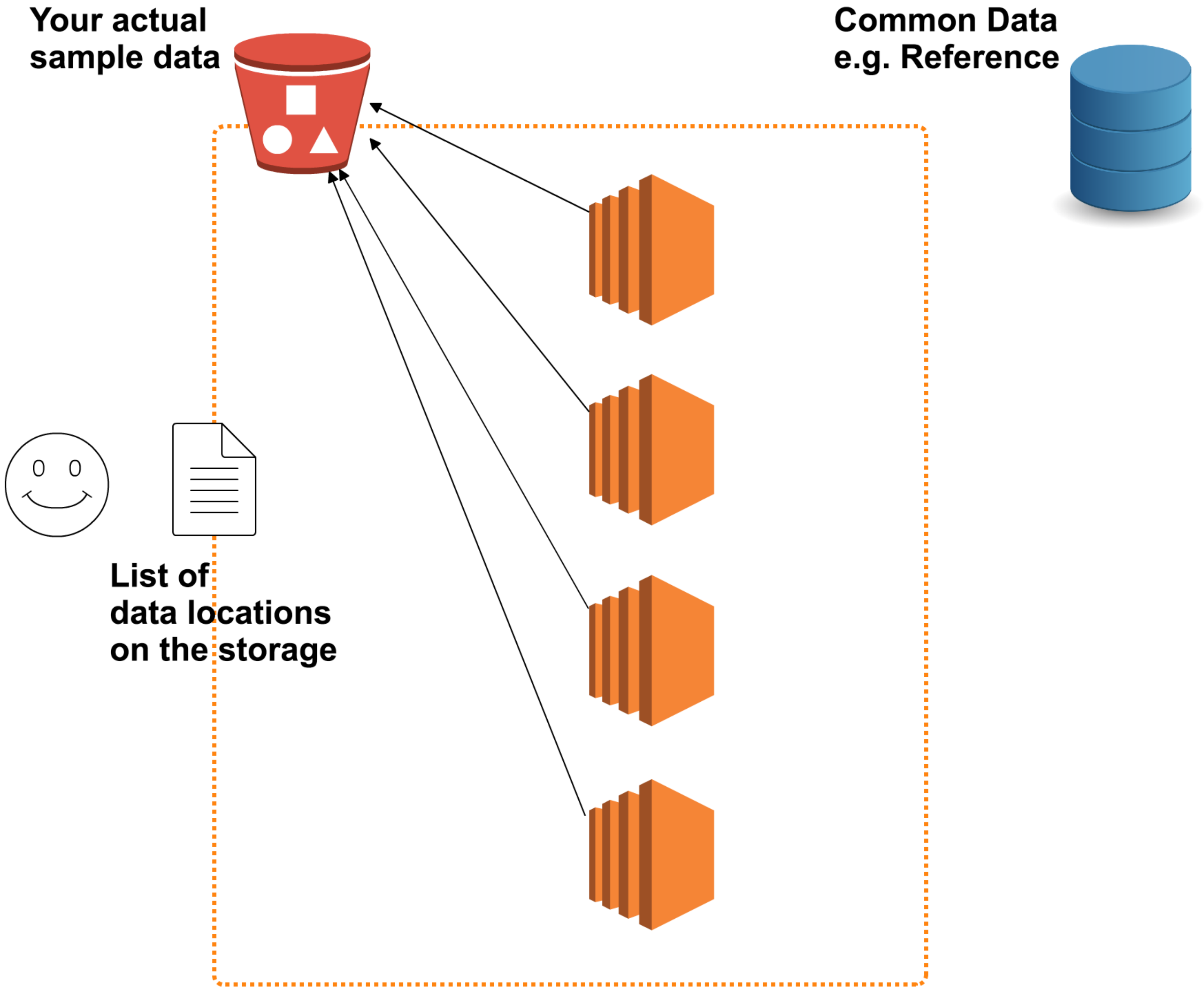


Your actual sample data

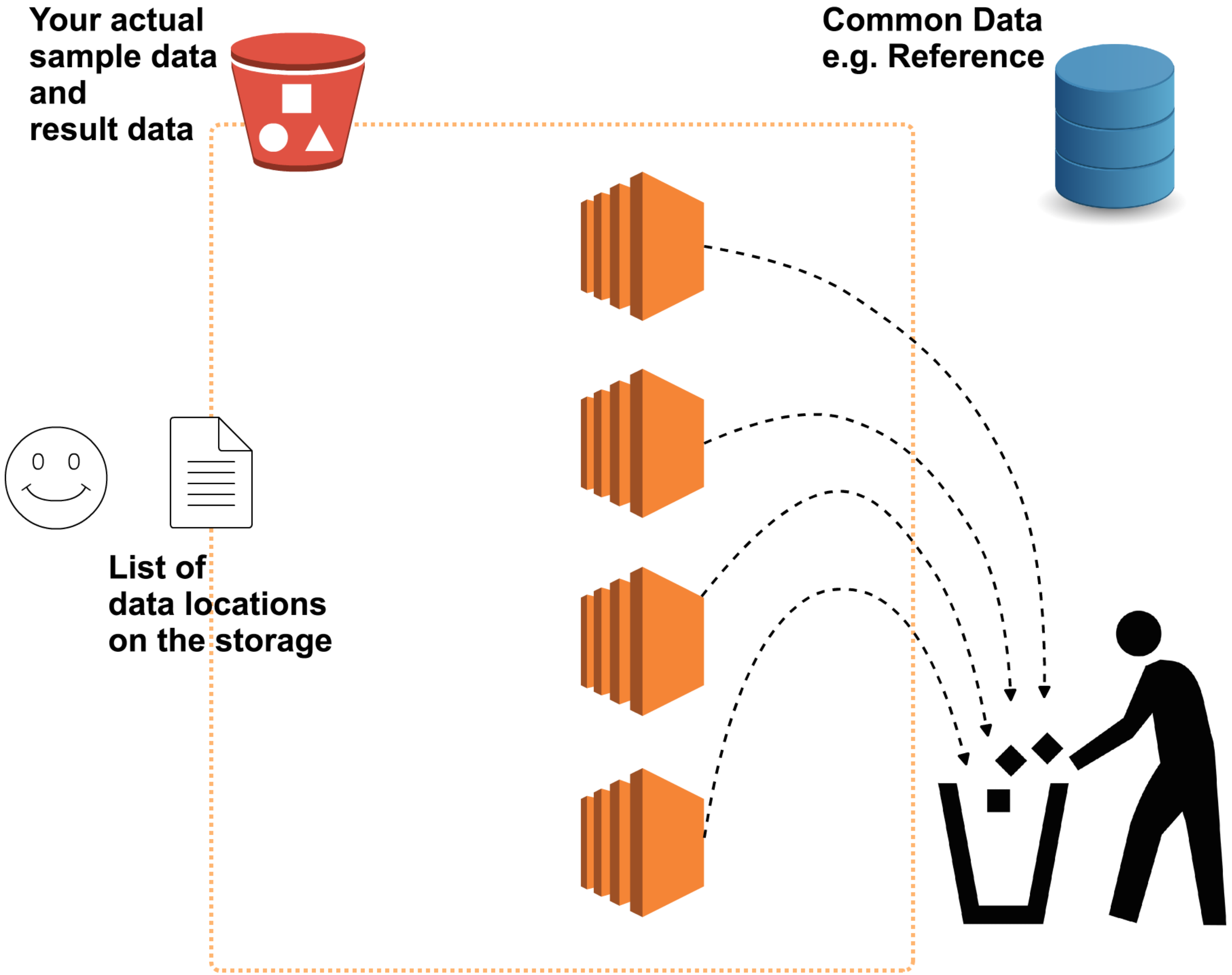
Common Data e.g. Reference

List of data locations on the storage

Push the result data back to the storage



Dispose all the computing resources no longer used



Your actual sample data and result data

Common Data e.g. Reference

List of data locations on the storage

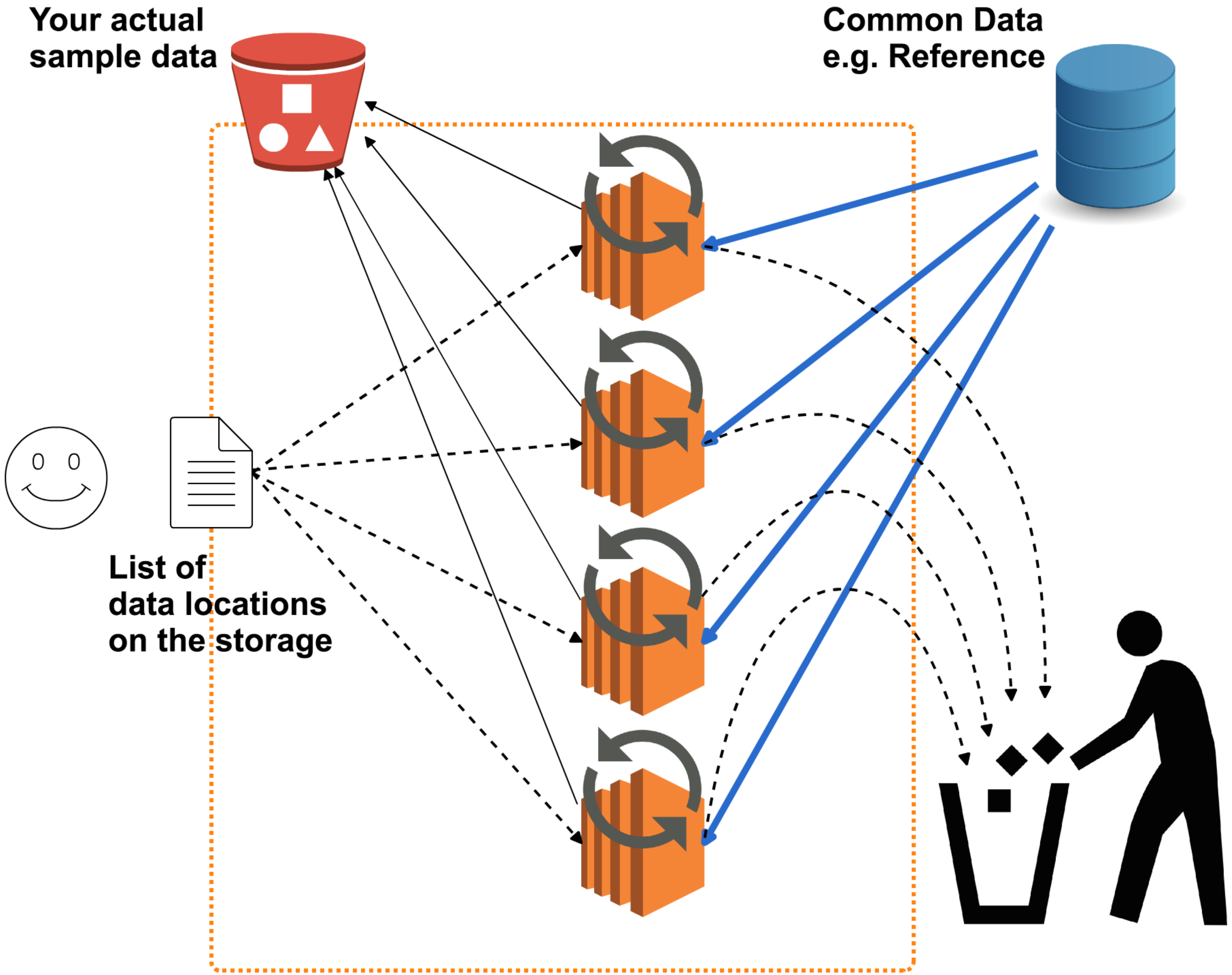
All you got is the result data!

Your actual
sample data
and
result data

Common Data
e.g. Reference

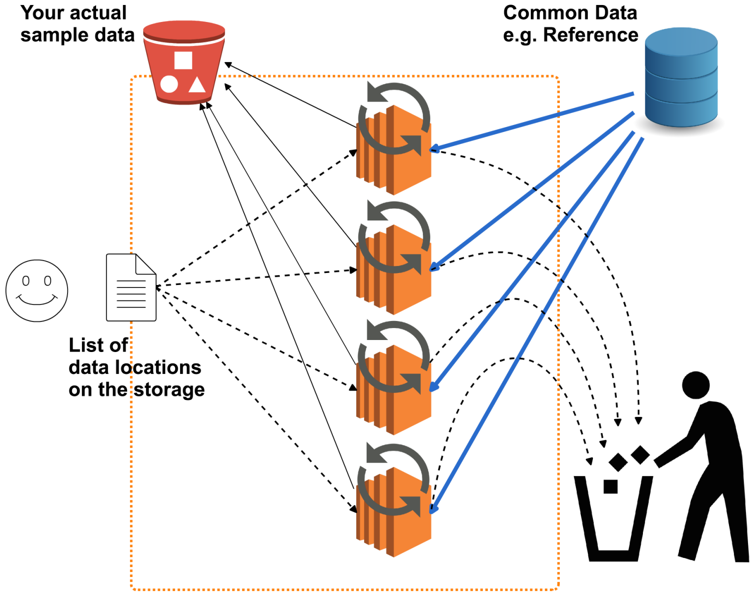
List of
data locations
on the storage

Overall


Your actual
sample data
Common Data
e.g. Reference
List of
data locations
on the storage

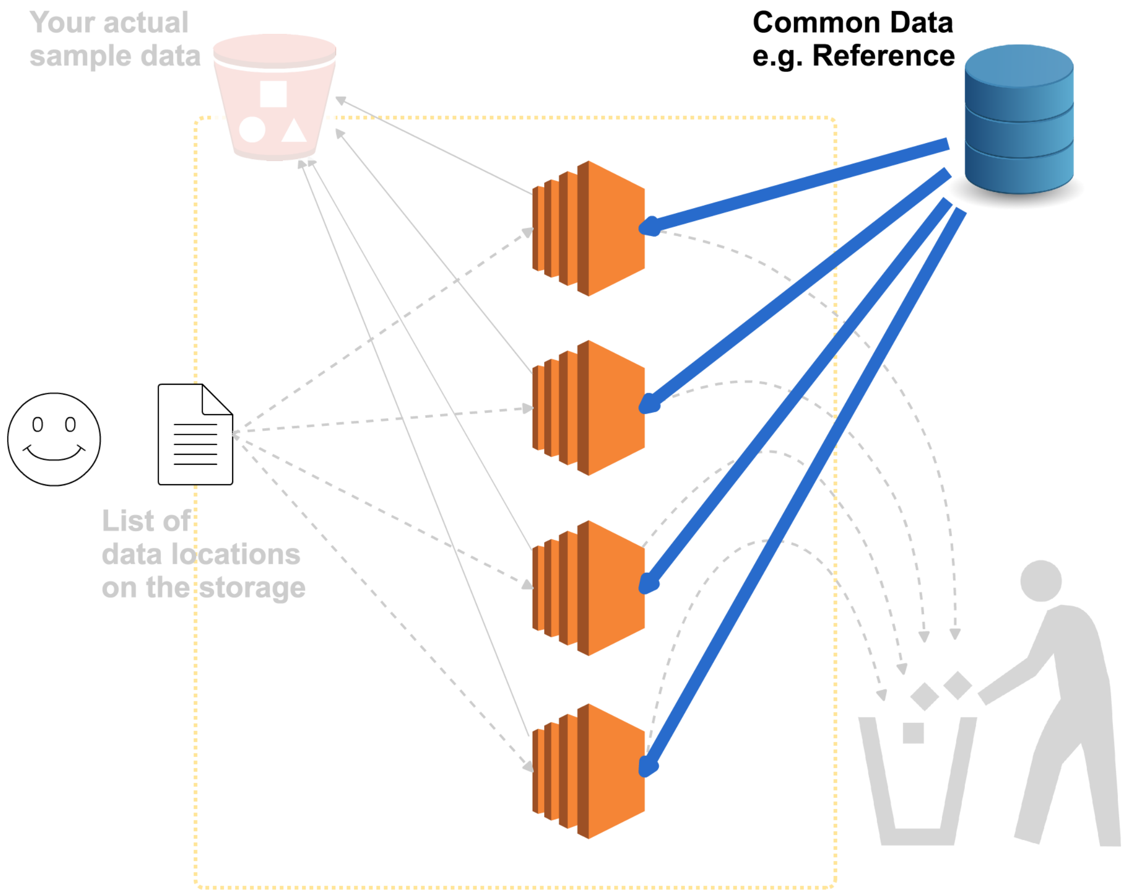
by using `awsub`

```
$ awsub \
  --tasks  ./my-samples.csv \
  --script ./my-workflow.sh \
  --image  otiai10/STAR-alignment # any Docker image
```


Your actual sample data
Common Data e.g. Reference
List of data locations on the storage

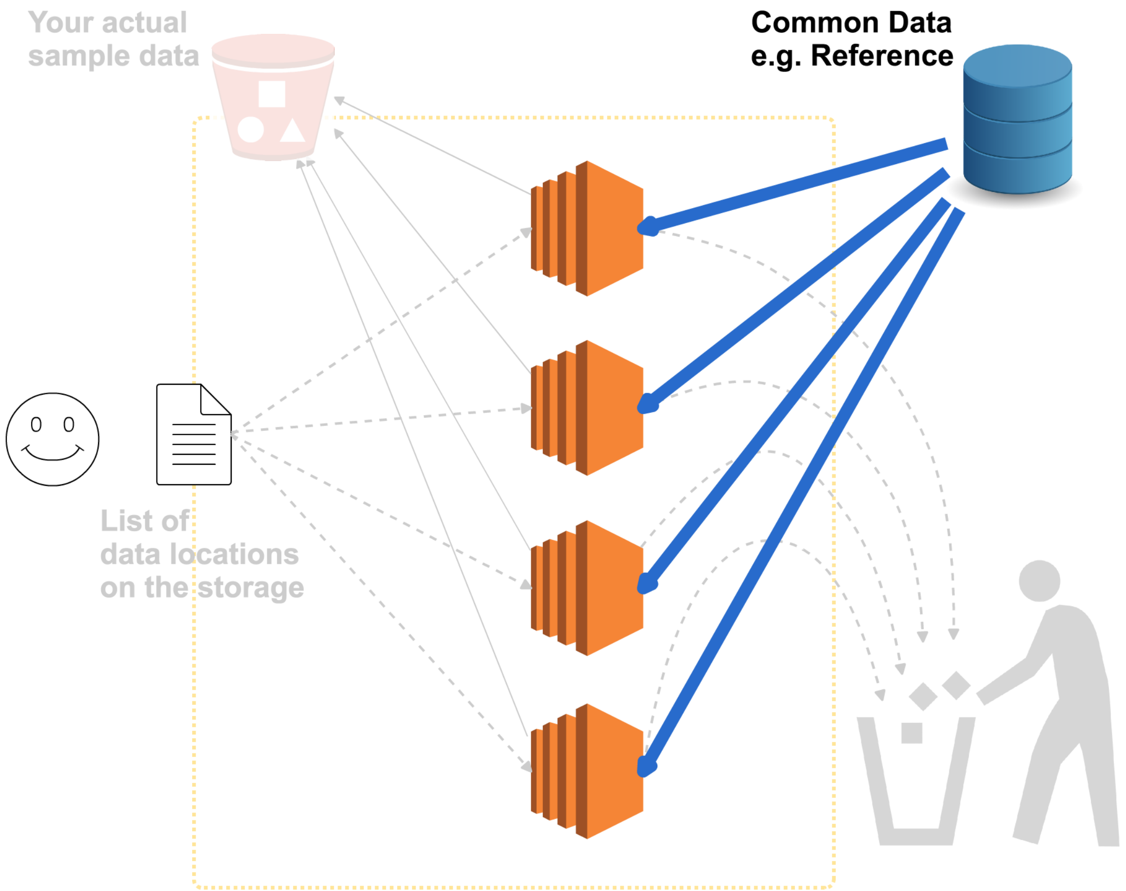
Problems of ETL on Bioinformatics

Problems of ETL on Bioinformatics


Your actual sample data
Common Data e.g. Reference
List of data locations on the storage

- Common Reference Data is so huge
    - Copying huge reference data uses
        - inefficient **traffic**
        - inefficient **instance time**
        - inefficient **storage area**
    - 具体的な例: ヒトのSTARで、40G弱

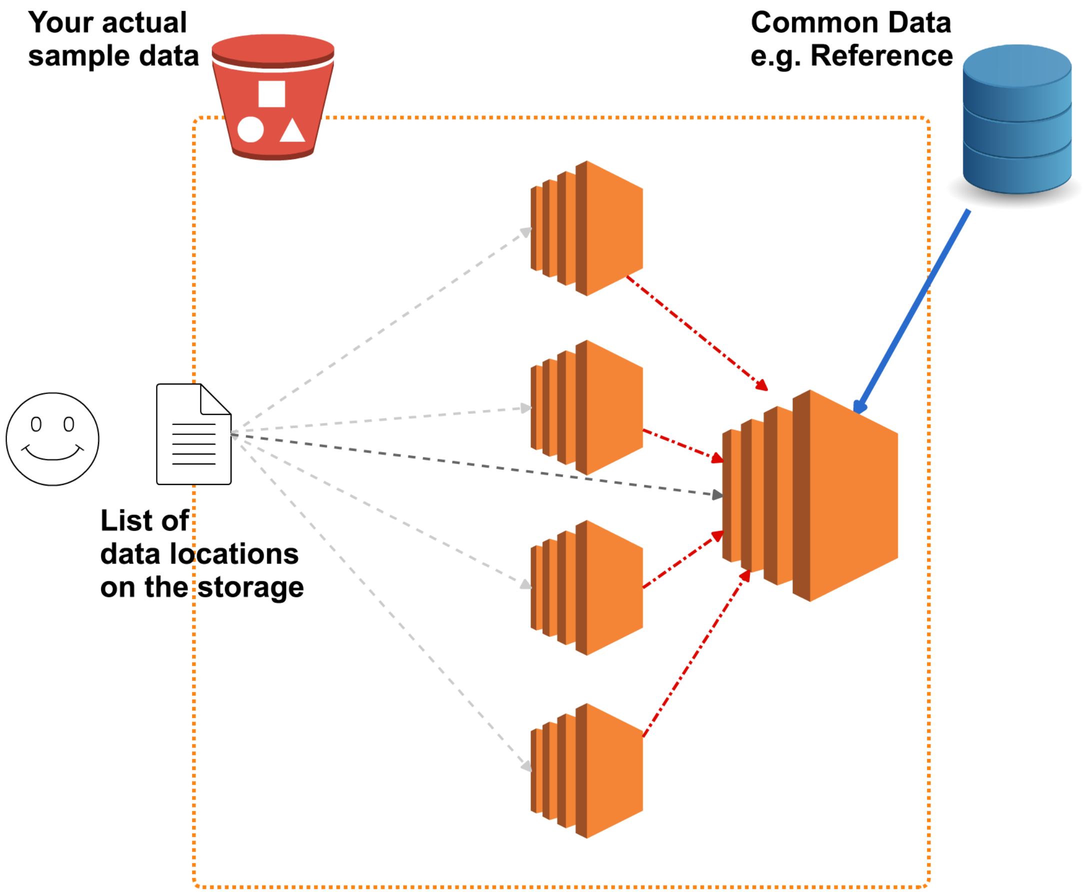
Problems of ETL on Bioinformatics



- Common Reference Data is so huge
    - Copying huge reference data uses
        - inefficient **traffic**
        - inefficient **instance time**
        - inefficient **storage area**
-

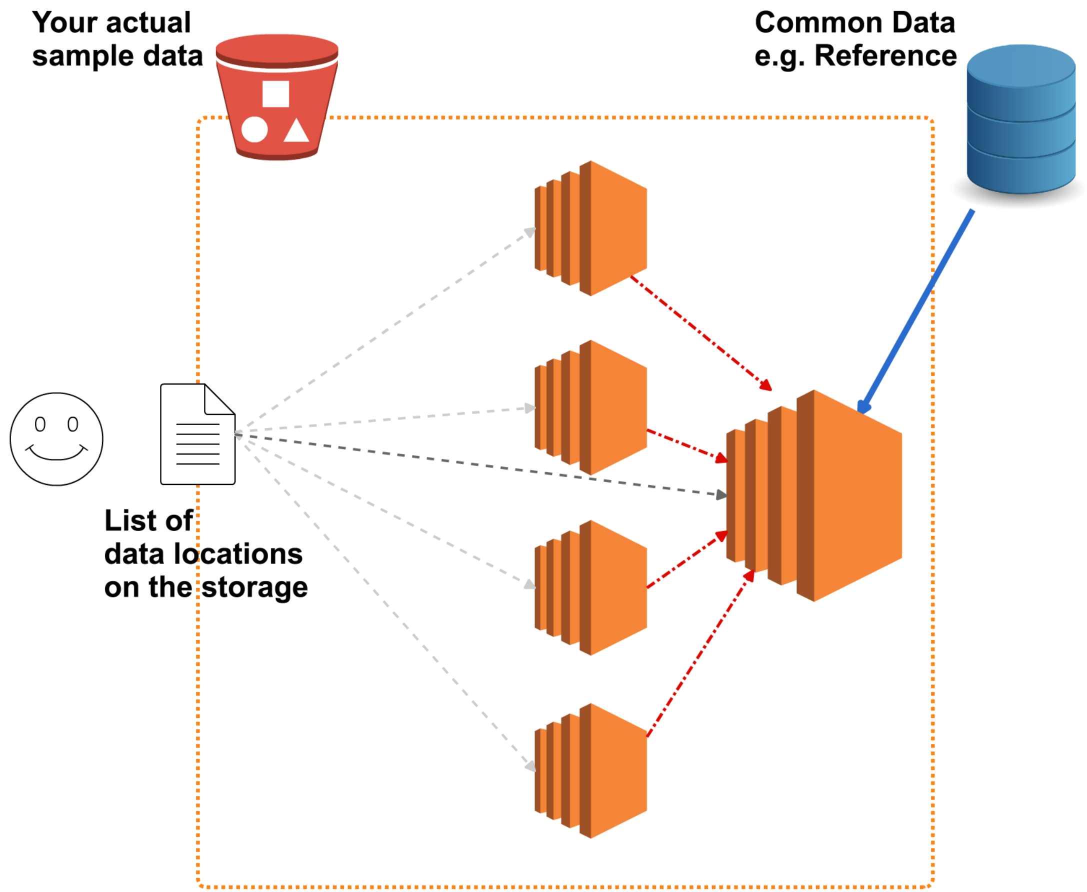
Suggestion: *Extended* ETL

Suggestion: *Extended* ETL



- Create a `Shared Data Instance`
- Fetch external common data **once**
- Let computing instances **mount**

Suggestion: *Extended* ETL



- Create a `Shared Data Instance`
- Fetch external common data **once**
- Let computing instances **mount**

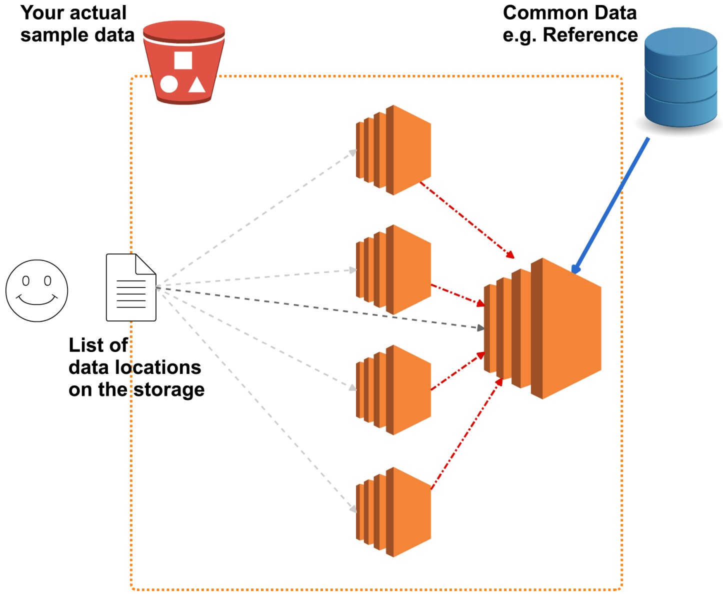
Cost Saving!

- Network traffic, instance time, ...

// ここにfigureを入れる

ExTL by using `awsub`

```
$ awsub \
  --tasks  ./my-samples.csv \
  --script ./my-workflow.sh \
  --image  otiai10/STAR-alignment \
+ --shared REFERENCE=s3://bucket/huge/reference
```



Summary

- Another approach than "Cluster on Cloud"
  - **"On-demand ETL on Cloud"**
- Huge common data can be a problem of "ETL on Cloud"
- **"Extended ETL"** (ExTL)
- Working Example Implementation of ExTL: `awsub`

More on the poster

about

- How to **Get started**
- **Google Cloud**, Microsoft Azure, OpenStack and more
- Common Workflow Language (**CWL**)
- Execution **Protocol** and Security Groups / IAM Instance Profile
- ***Go*** implementation
- etc...

Come to poster **B29**, and any feedback is welcome!

https://github.com/otiai10/awsub