awsub

An Extended ETL engine for cloud computing resources.

Hiromu OCHIAI - National Cancer Center Japan

Cluster

- Not everyone can use HPC
- It's a hard work to build and maintain clusters on cloud

ETL on Cloud

- Extract, Transform, Load
- On-demand resource procurement
- for example
    - AWS batch, dsub, etc...

`awsub` as ETL engine

```
% awsub \
  --image otiai10/STAR \
  --script ./my-workflow.sh \
  --tasks ./samples.csv \
  --aws-instance-type m4.2xlarge
```

// 図

Problems of ETL on Bioinfo

- Big common data, e.g. Reference file
  - Network cost
  - Instance cost

// 図

Suggestion: Extended ETL

Extended ETL data processing model

// 図

Implementation of ExTL

```
% awsub \
  --image otiai10/STAR \
  --script ./my-workflow.sh \
  --tasks ./samples.csv \
  --aws-instance-type m4.2xlarge \
+ --shared REFERENCE=s3://bucket/path/to/reference
```

More on the poster

about...

- Google Cloud, Microsoft Azure, OpenStack and more
- Common Workflow Language (CWL)
- Execution Protocol and Security Groups
- etc

Come to poster **#55**, and any feedback is welcome!

https://github.com/otiai10/awsub