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
```

// figure here

Problems of ETL on Bioinfo

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

Suggestion: Extended ETL

Extended ETL data processing model

// figure here

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
```

// figure here