How to Use Azure Batch Service For Variant Loading

OpenCGA loads a huge number of variants files (Terabyte to Petabyte in size) into storage and is one of its critical and time consuming task. OpenCGA striving hard to reduce the time from availability of file to enable user to query it using OpenCGA. OpenCGA latest version (1.4.0 rc3) enables users to use "Azure Batch Service" which can large scale out and thus distribute load and index variant files in parallel. This will significantly increase the variant loading time and OpenCGA performance.

OpenCGA uses ARM template to auto deploy a pool with preconfigured opencga docker image. This pool is *"AutoScale"* enabled so will scale as number of variant index jobs will grow. These ARM script will also populate the following section in *"configuration.yml"* which enables OpenCGA daemon to prepare catalog job and then submit it as azure task to Azure Batch Service.

Configuration

configuration.yml

```
....
execution:
  mode: AZURE
        ...
  options:
    #Azure Batch Service information
    batchAccount : "batchAccount"
    batchKey : "batchKey"
    batchUri : "https://batchservice.uksouth.batch.azure.com"
    batchPoolId : "poolId"
    dockerImageName : "openCGADockerImageName" # preconfigured docker image
    dockerArgs : "dockerRunOptions"   # e,g; mount points etc.
....
```

Variant Index Job Creation

Once user create an OpenCGA variant indexing job, this will be stored in OpenCGA catalog. For example, following is an example to link a file in catalog and then create index pipeline which internally will be stored as a catalog job :

OpenCGA Variant Index Job Creation

```
./opencga.sh files link -i variantFile.vcf.gz -s myStudy
./opencga.sh variant index --file variantFile.vcf.gz --calculate-stats --annotate -o tmp
```

If OpenCGA daemon is not running, user can start it with the following command:

OpenCGA Daemon

```
/opt/opencga/bin$ ./opencga-admin.sh catalog daemon --start <<< admin_password
```

Once daemon is running, it will fetch available jobs from the catalog, prepare them and then submit each catalog job as an "*Azure Task*" to the batch pool specified in *"configuration.yml"*. A typical Azure task command will look like :

Azure Batch Service Task

```
/opt/opencga/bin/opencga-analysis.sh variant index --outdir /opt/opencga/sessions/jobs/J_2510 -
DcalculateStats=true --annotate --file variantFile.vcf.gz --path tmp: ...
```

On startup, docker container will mount the locations listed in *"dockerArgs"* parameter in *"configuration.yml"* file, e.g; *"/opt/opencga/conf"*, *"/opt/opencga /sessions", "storage location"* (where variant files are stored*)* and any other run time options. This docker container will have access to shared configuration, session and storage location and will start indexing the variant file into storage (HBase|MongoDB) as described in index pipeline.