De-identify an engineering corpus

Build your own pipeline to replace identifying information with **consistent pseudonyms**, while preserving usable engineering files and the links between them. The input is raw: its original identifiers are intentionally present.

Assessment: each submitted attempt receives a score out of **1,000**. You may submit **up to three attempts**, improving your pipeline between attempts. Include your pipeline's source code with every attempt.

1. Copy the input into a bucket you own

The source contains approximately 139.7 GB across 188,722 files, including PDFs, drawings, SolidWorks and Inventor files, neutral CAD formats, images and archives. The AWS region is `us-west-2`. Source:

```
s3://machina-deid-v1-challenge/releases/v1/corpus/
```

Use AWS CLI v2 and an AWS profile authenticated to **your own account**. Replace the two values below. The bucket name must be globally unique. These commands create a fresh bucket and copy the raw input into its private `input/` folder. Your account pays its AWS storage and request charges.

```
PROFILE=candidate
BUCKET=your-unique-deid-bucket

aws s3api create-bucket --bucket "$BUCKET" \
  --profile "$PROFILE" --region us-west-2 \
  --create-bucket-configuration LocationConstraint=us-west-2 \
  --object-ownership BucketOwnerEnforced

aws s3 sync \
  s3://machina-deid-v1-challenge/releases/v1/corpus/ \
  "s3://$BUCKET/input/" \
  --profile "$PROFILE" --source-region us-west-2 \
  --region us-west-2 --copy-props none --sse AES256
```

Use your credentials for the destination write; do not add `-no-sign-request` to this copy command. Keep raw input private.

2. Write and run your de-identification pipeline

Write your own reproducible pipeline to process the full collection. Run it using your preferred infrastructure, with the copied input as its source. Keep temporary files and private mappings outside the submission folder. If downloading locally, use a case-sensitive filesystem and allow space for expanded archives and output.

What the pipeline must achieve

- Replace identifying company and person names, accounts, contact details, domains, addresses, logos, and identifying project or part names and numbers. The same original identity must receive the same pseudonym throughout the corpus; different identities must remain distinct. Preserve meaningful pseudonyms in fields where an identity belongs, rather than simply blanking those fields.
- Cover visible text and images, filenames and folder names, PDF metadata and attachments, native CAD properties, embedded previews, and archive contents. Inspect internal metadata in native files where applicable; a plain text search or a clean preview is not sufficient.
- Preserve geometry, dimensions, tolerances, units, materials, configurations, BOMs, and relationships between drawings, models and assemblies. Update renamed files and their references together. Keep usable native files, not only exported PDFs or images. Do not alter format-required keys or non-identifying engineering information.

The input includes duplicates and some pre-existing invalid or empty files. Do not silently discard difficult engineering files. Explain source defects and any omissions in your submission notes, using final paths or neutral descriptions. Original identifiers must not remain in unprocessed copies inside output archives.

3. Upload the cleaned corpus and your source code

Create a separate folder for each attempt in **your own submission bucket**:

```
input/                          private raw input
submissions/attempt-1/
  corpus/                       de-identified files and archives
  code/                         pipeline source and dependencies
  README.txt                    how to run; limitations and exceptions
```

Include the complete source used for that attempt, dependency versions or lockfiles, and exact run instructions. The source may contain identifier-matching logic; the de-identification score evaluates `corpus/`. Do not include credentials, raw input data, temporary files, or private run logs in the public submission.

Prepare local folders `cleaned-corpus/` and `pipeline-source/`, plus `README.txt`. Then upload them. Keep the same `PROFILE` and `BUCKET` values from page 1.

```
ATTEMPT=1
DEST="s3://$BUCKET/submissions/attempt-$ATTEMPT"

aws s3 sync ./cleaned-corpus/ "$DEST/corpus/" \
  --profile "$PROFILE" --region us-west-2 \
  --sse AES256 --no-follow-symlinks
aws s3 sync ./pipeline-source/ "$DEST/code/" \
  --profile "$PROFILE" --region us-west-2 \
  --sse AES256 --no-follow-symlinks
aws s3 cp ./README.txt "$DEST/README.txt" \
  --profile "$PROFILE" --region us-west-2 --sse AES256
```

For attempts 2 and 3, change `ATTEMPT` and upload the full output and code again. Once you submit an attempt, leave its contents unchanged during grading.

4. Make your submission readable and send its location

For the fresh bucket created on page 1, the following policy allows anyone to read and list `submissions/`, including the submitted code. It grants no public writes and does not expose `input/`. It will cover all three attempts. Run this once, after reviewing the contents you are publishing.

```
cat > submission-policy.json <<EOF
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow", "Principal": "*",
      "Action": "s3:GetObject",
      "Resource": "arn:aws:s3:::$BUCKET/submissions/*"
    },
    {
      "Effect": "Allow", "Principal": "*",
      "Action": "s3:ListBucket",
      "Resource": "arn:aws:s3:::$BUCKET",
      "Condition": {"StringLike": {
        "s3:prefix": ["submissions/", "submissions/*"]
      }}
    }
  ]
}
EOF

aws s3api put-public-access-block --bucket "$BUCKET" \
  --profile "$PROFILE" --region us-west-2 \
  --public-access-block-configuration \
  '{"BlockPublicAcls":true,"IgnorePublicAcls":true,
    "BlockPublicPolicy":false,"RestrictPublicBuckets":false}'
aws s3api put-bucket-policy --bucket "$BUCKET" \
  --profile "$PROFILE" --region us-west-2 \
  --policy file://submission-policy.json
```

These commands assume a new bucket with no existing policy. If your account or organization blocks public buckets, contact the organizer to arrange access; do not change account-wide controls or share AWS credentials.

Check access without credentials:

```
aws s3 ls "$DEST/" --recursive \
  --no-sign-request --region us-west-2
aws s3 cp "$DEST/README.txt" - \
  --no-sign-request --region us-west-2
```

Also test downloading a file from `corpus/` and one from `code/`. Send the organizer your name or candidate ID, **attempt number**, region (`us-west-2`), and the complete submission location:

```
s3://your-unique-deid-bucket/submissions/attempt-1/
```

You will receive that attempt's score out of 1,000 and grading feedback. Use the remaining attempts, up to three total, to improve and resubmit your pipeline. Keep each submitted attempt available until the organizer confirms grading is complete.