Council on Dairy Cattle Breeding

Data Ingestion System

Ezequiel L. Nicolazzi, CDCB Chief Operating Officer

CDCB

COUNCIL ON DAIRY CATTLE BREEDING

CDCB data ingestions systems

- **All** records that you submit are processed by a group of programs that perform the “ETL” process: extract, transform, load.
- The 2 main programs are named:
  - **EDCLOB** (genomic data ingestion system)
  - **EDITS** (pretty much everything else you submit, including nominations)
    - The *true* data ingestion system
- Runs and writes locally to the database.
- Active 24/7

CDCB data ingestions systems – What we're changing

- CDCB is investing resources in modernizing its infrastructure.
  - Is there something wrong with EDITS?
  - Create more opportunities for the US industry
  - Use more modern programming infrastructure and languages
  - Full documentation of business rules
  - Succession plan
- WebConnect was the first of such projects.
- **CDCB data ingestion systems are next.**
  - **EDITS is phase 1.**

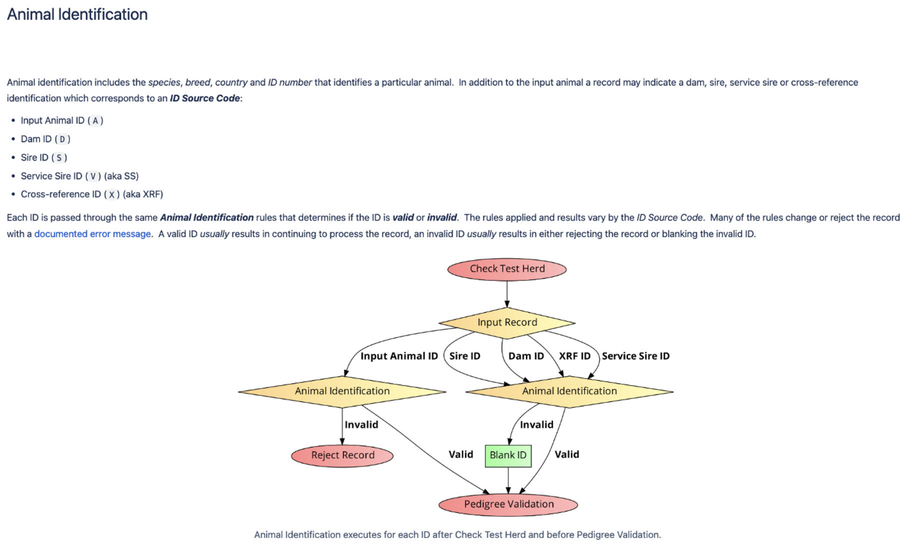
CDCB data ingestions system project

- The project concept was developed way before it was even conceived.
- First **key** step: full documentation of business rules
- Thousands of lines of documentation.. inside the code.
    - … and in the head of a CDCB employee.

Documentation of business rules

iYOTAH SOLUTIONS

- Formal documentation in Confluence
- Collaboration with iYotah solutions
- ~ 1 year for one highly skilled programmer.
- Independent review of business rules
  - very little # of discarded rules
- Reality check.



Animal Identification

Animal identification includes the *species*, *breed*, *country* and *ID number* that identifies a particular animal. In addition to the input animal a record may indicate a dam, sire, service sire or cross-reference identification which corresponds to an ***ID Source Code***:

- Input Animal ID (`A`)
- Dam ID (`D`)
- Sire ID (`S`)
- Service Sire ID (`V`) (aka SS)
- Cross-reference ID (`X`) (aka XRF)

Each ID is passed through the same ***Animal Identification*** rules that determines if the ID is ***valid*** or ***invalid***. The rules applied and results vary by the *ID Source Code*. Many of the rules change or reject the record with a documented error message. A valid ID *usually* results in continuing to process the record, an invalid ID *usually* results in either rejecting the record or blanking the invalid ID.

Animal Identification executes for each ID after Check Test Herd and before Pedigree Validation.

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

The next phase: planning the future

- There is nothing wrong with the current system
- But there are limitations:
  - Local server (pros and cons!)
  - Physical storage space
  - A single CDCB staff member in control of development
    - Very difficult skills to pass along
  - Tightly linked to structured DHI data
  - Not possible to integrate to new tools developed or in development

Planning the future

- CDCB is expected to:
  - Provide solutions at a faster rate than in the past
  - Accept a wider range of data formats and providers
  - Deal with (and store) bigger data (e.g. sensors, MIR)
  - Provide better feedback to users (and help them solve issues)
  - Store large amounts of raw data for researchers to make sense of
  - Etc…
- The original Data Ingestion System was not designed to tackle these new needs

Designing the future ...

- Re-write the Data Ingestion System (~ no change of rules) to enhance its functionalities:
    - ID database
    - File queuing system
    - New universal format for data submission
    - API capabilities
    - Cloud-based system
    - Unlimited storage
    - New reporting capabilities & direct integration with WebConnect

… without disrupting operations

- Biggest challenge of all
- Interaction with our current (local) database
- Minor / no changes to user interaction necessary
- Mixed system: local + cloud interaction needs to be seamless

Developing the future

- A year and a half project signed with iYotah solutions
- Exclusive team of 6 people dedicated to this project
- 5 deliverables
- Objective: reproduce EDITS on a cloud environment successfully interacting with CDCB system (full integration).
    - Barely any change for users of the *current* system in the delivery of v.1.0
    - Further enhancements already spec'd, planned, and documented

What “barely any change” means

- No change for the user, completely new system “under the hood”
- Maintaining rules and functionalities will allow thorough testing and guarantee seamless continuity of services
- Some new features:
  - possibility of submitting files via API
  - some basic statistics shown in WebConnect
  - “visible” queuing system (where is my file in the queue?)

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Where are we in the project

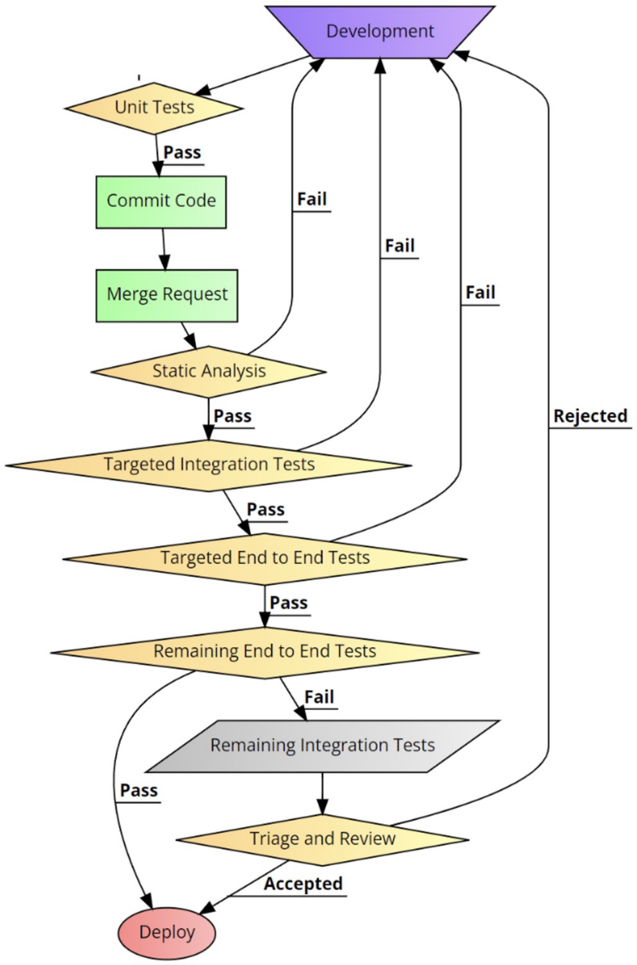
- **Deliverable 1**: “Cloud + Data lake + infrastructure setup” (Q2 2023)
- **Deliverable 2**: “Universal format, API” (Q3 2023)
- **Deliverable 3**: “Testing framework, ID database” (Q4 2023)
- **Deliverable 4**: “EDITS code, integration testing” (Q2 2024) – Expected Q3 2024
- **Deliverable 5**: “UAT” (Q3 2024) – Expected Q4 2024

How testing will work

- The project includes 4 phases of testing:
    - iYotah’s testing platform (cucumber): automated set of rules continuously maintained to ensure future changes to the code do not affect other parts of the system.
    - CDCB targeted testing: CDCB staff is testing features as they are released
    - CDCB integrated testing: CDCB staff will test integration of the new system vs the old by running both in parallel and compare output
    - UAT: Selected industry partners will test the system before launch

And after the project?

- Similar process to development, but faster
- *Continuous testing*
- Human (CDCB staff) approval still req'd!



COUNCIL ON DAIRY CATTLE BREEDING

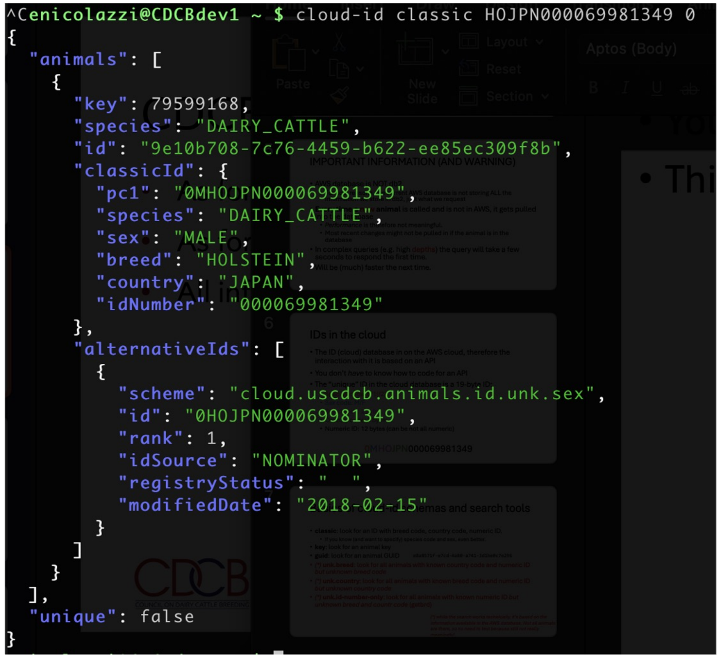
Setting clear expectations

- Release 1.0 will be a slightly enhanced version of the current system.
    - Maximum success: nothing changes for you
    - We *must* ensure the system is stable and robust before building new features
    - New features will be easier to implement
        - Much more flexibility
        - Automated testing infrastructure assistance
        - Can work in “islands” and deploy as many testing environments as we need
- Future releases:
    - Harvesting ideas internally first, will interact with industry next.
    - Nearly perfect alignment with feedback from 2023 CDCB Industry Workshop
    - Need to shape the future

SOME “VISUAL” RESULTS

CDCB ID database

- As for the user, there is no change to current practices.
- As for CDCB, this is planned to be the core of animal identification.
- All information linked to an animal, gathered centrally in a database.

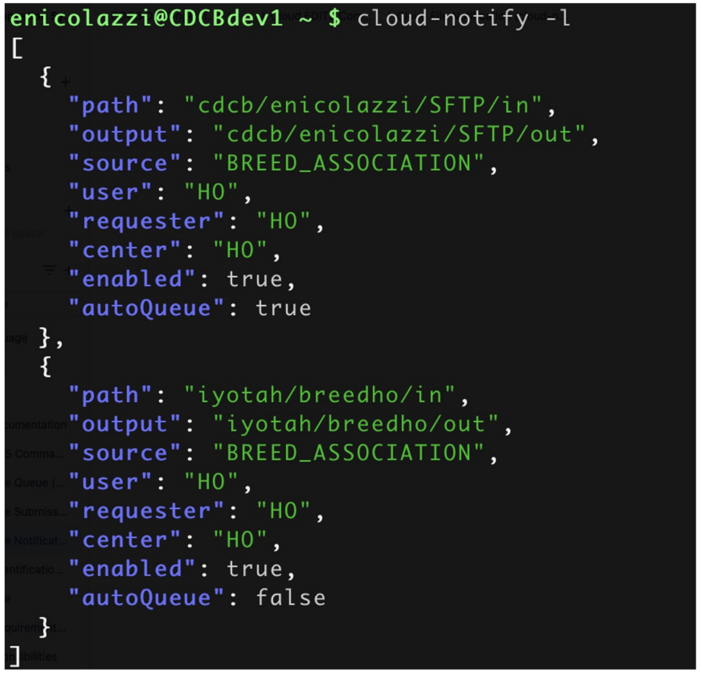


```
^Cenicolazzi@CDCBdev1 ~ $ cloud-id classic HOJPN000069981349 0
{
  "animals": [
    {
      "key": 79599168,
      "species": "DAIRY_CATTLE",
      "id": "9e10b708-7c76-4459-b622-ee85ec309f8b",
      "classicId": {
        "pc1": "0MHOJPN000069981349",
        "species": "DAIRY_CATTLE",
        "sex": "MALE",
        "breed": "HOLSTEIN",
        "country": "JAPAN",
        "idNumber": "000069981349"
      },
      "alternativeIds": [
        {
          "scheme": "cloud.uscdcb.animals.id.unk.sex",
          "id": "0HOJPN000069981349",
          "rank": 1,
          "idSource": "NOMINATOR",
          "registryStatus": "  ",
          "modifiedDate": "2018-02-15"
        }
      ]
    }
  ],
  "unique": false
}
```

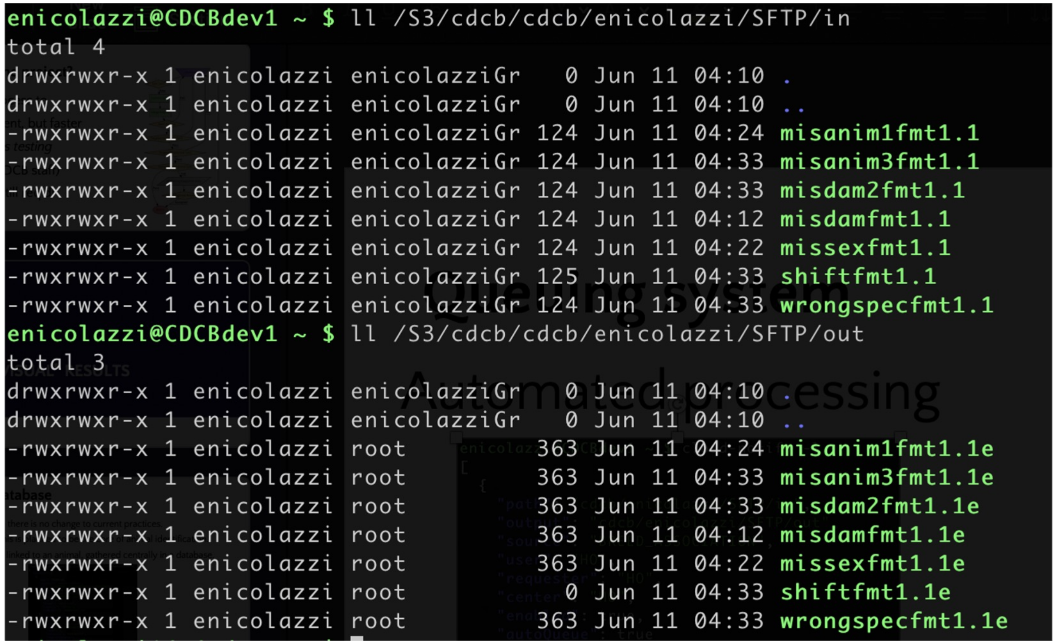
Queue system

- Replaces bash automations handling files from your “in” folders to EDITS and back to your “out” folders
  - (CDCB only) Processing a file from.. anywhere
  - (CDCB only) Processing a file from a cloud location
  - (all) Automated processing of a file in a folder (e.g. SFTP)
  - (all) API submission to the queue

Queue system – Automated processing

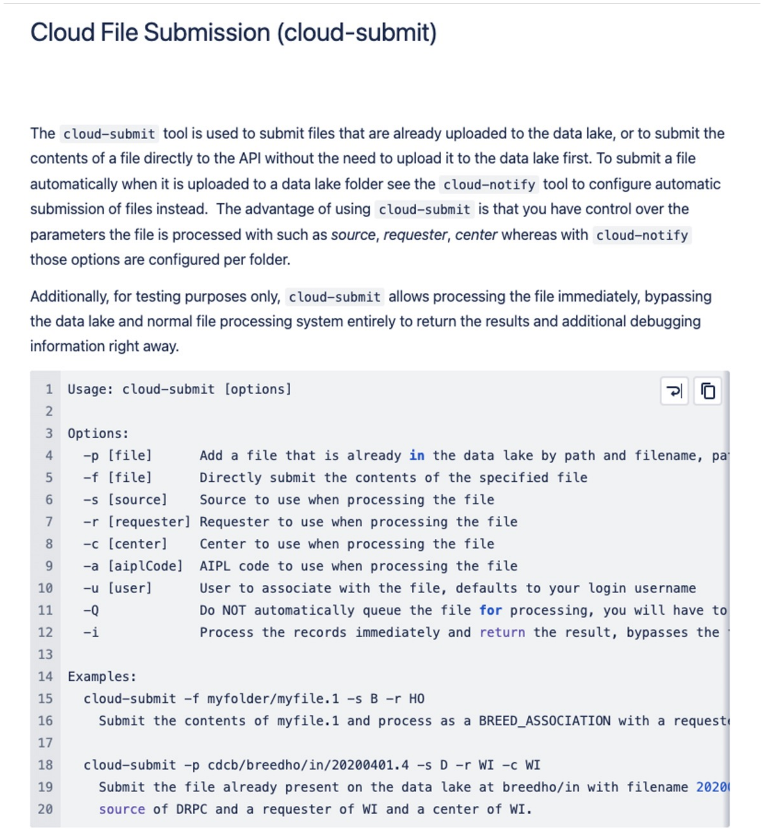


```
enicolazzi@CDCBdev1 ~ $ cloud-notify -l
[
  {
    "path": "cdcb/enicolazzi/SFTP/in",
    "output": "cdcb/enicolazzi/SFTP/out",
    "source": "BREED_ASSOCIATION",
    "user": "HO",
    "requester": "HO",
    "center": "HO",
    "enabled": true,
    "autoQueue": true
  },
  {
    "path": "iyotah/breedho/in",
    "output": "iyotah/breedho/out",
    "source": "BREED_ASSOCIATION",
    "user": "HO",
    "requester": "HO",
    "center": "HO",
    "enabled": true,
    "autoQueue": false
  }
]
```



```
enicolazzi@CDCBdev1 ~ $ ll /S3/cdcb/cdcb/enicolazzi/SFTP/in
total 4
drwxrwxr-x 1 enicolazzi enicolazziGr   0 Jun 11 04:10 .
drwxrwxr-x 1 enicolazzi enicolazziGr   0 Jun 11 04:10 ..
-rwxrwxr-x 1 enicolazzi enicolazziGr 124 Jun 11 04:24 misanim1fmt1.1
-rwxrwxr-x 1 enicolazzi enicolazziGr 124 Jun 11 04:33 misanim3fmt1.1
-rwxrwxr-x 1 enicolazzi enicolazziGr 124 Jun 11 04:33 misdam2fmt1.1
-rwxrwxr-x 1 enicolazzi enicolazziGr 124 Jun 11 04:12 misdamfmt1.1
-rwxrwxr-x 1 enicolazzi enicolazziGr 124 Jun 11 04:22 missexfmt1.1
-rwxrwxr-x 1 enicolazzi enicolazziGr 125 Jun 11 04:33 shiftfmt1.1
-rwxrwxr-x 1 enicolazzi enicolazziGr 124 Jun 11 04:33 wrongspecfmt1.1
enicolazzi@CDCBdev1 ~ $ ll /S3/cdcb/cdcb/enicolazzi/SFTP/out
total 3
drwxrwxr-x 1 enicolazzi enicolazziGr   0 Jun 11 04:10 .
drwxrwxr-x 1 enicolazzi enicolazziGr   0 Jun 11 04:10 ..
-rwxrwxr-x 1 enicolazzi root         363 Jun 11 04:24 misanim1fmt1.1e
-rwxrwxr-x 1 enicolazzi root         363 Jun 11 04:33 misanim3fmt1.1e
-rwxrwxr-x 1 enicolazzi root         363 Jun 11 04:33 misdam2fmt1.1e
-rwxrwxr-x 1 enicolazzi root         363 Jun 11 04:12 misdamfmt1.1e
-rwxrwxr-x 1 enicolazzi root         363 Jun 11 04:22 missexfmt1.1e
-rwxrwxr-x 1 enicolazzi root           0 Jun 11 04:33 shiftfmt1.1e
-rwxrwxr-x 1 enicolazzi root         363 Jun 11 04:33 wrongspecfmt1.1e
```

Queue system – CLOUD / API submission



Cloud File Submission (cloud-submit)

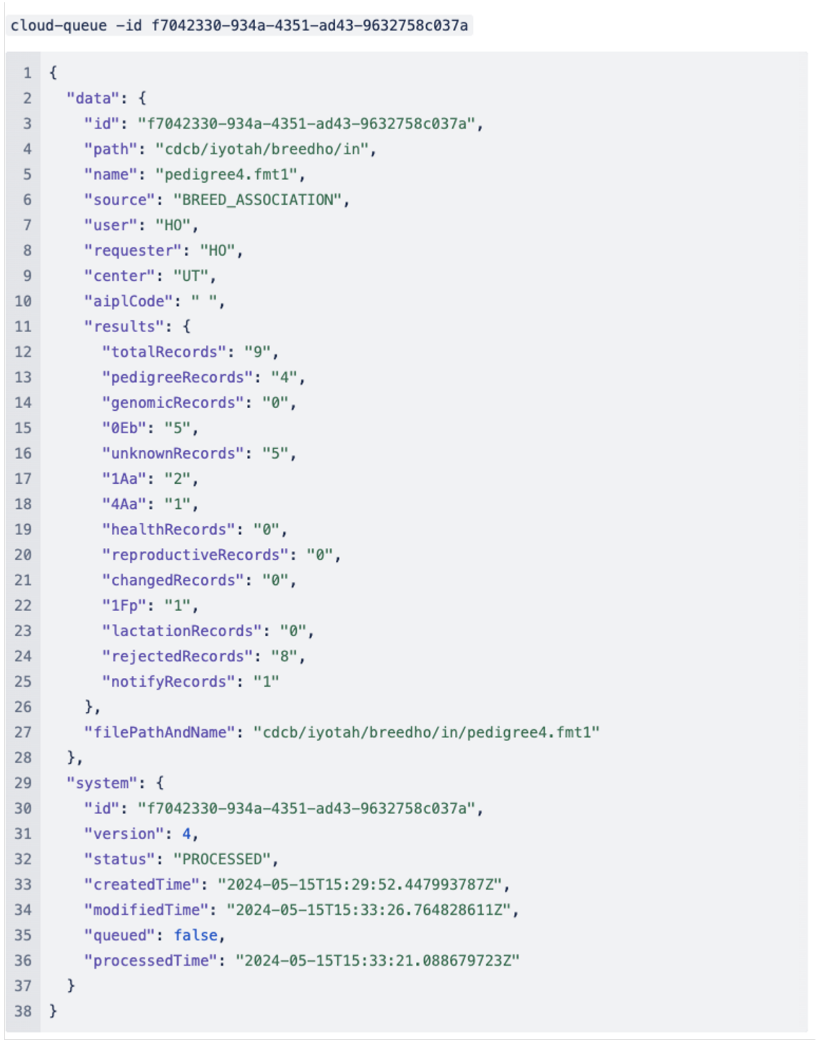
The `cloud-submit` tool is used to submit files that are already uploaded to the data lake, or to submit the contents of a file directly to the API without the need to upload it to the data lake first. To submit a file automatically when it is uploaded to a data lake folder see the `cloud-notify` tool to configure automatic submission of files instead. The advantage of using `cloud-submit` is that you have control over the parameters the file is processed with such as *source*, *requester*, *center* whereas with `cloud-notify` those options are configured per folder.

Additionally, for testing purposes only, `cloud-submit` allows processing the file immediately, bypassing the data lake and normal file processing system entirely to return the results and additional debugging information right away.

```
Usage: cloud-submit [options]

Options:
  -p [file]      Add a file that is already in the data lake by path and filename, pa
  -f [file]      Directly submit the contents of the specified file
  -s [source]    Source to use when processing the file
  -r [requester] Requester to use when processing the file
  -c [center]    Center to use when processing the file
  -a [aiplCode]  AIPL code to use when processing the file
  -u [user]      User to associate with the file, defaults to your login username
  -Q             Do NOT automatically queue the file for processing, you will have to
  -i             Process the records immediately and return the result, bypasses the

Examples:
  cloud-submit -f myfolder/myfile.1 -s B -r HO
    Submit the contents of myfile.1 and process as a BREED_ASSOCIATION with a requeste

  cloud-submit -p cdcb/breedho/in/20200401.4 -s D -r WI -c WI
    Submit the file already present on the data lake at breedho/in with filename 2020
    source of DRPC and a requester of WI and a center of WI.
```

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Initial processing



```
cloud-queue -id f7042330-934a-4351-ad43-9632758c037a
```

```
{
  "data": {
    "id": "f7042330-934a-4351-ad43-9632758c037a",
    "path": "cdcb/iyotah/breedho/in",
    "name": "pedigree4.fmt1",
    "source": "BREED_ASSOCIATION",
    "user": "HO",
    "requester": "HO",
    "center": "UT",
    "aiplCode": " ",
    "results": {
      "totalRecords": "9",
      "pedigreeRecords": "4",
      "genomicRecords": "0",
      "0Eb": "5",
      "unknownRecords": "5",
      "1Aa": "2",
      "4Aa": "1",
      "healthRecords": "0",
      "reproductiveRecords": "0",
      "changedRecords": "0",
      "1Fp": "1",
      "lactationRecords": "0",
      "rejectedRecords": "8",
      "notifyRecords": "1"
    },
    "filePathAndName": "cdcb/iyotah/breedho/in/pedigree4.fmt1"
  },
  "system": {
    "id": "f7042330-934a-4351-ad43-9632758c037a",
    "version": 4,
    "status": "PROCESSED",
    "createdTime": "2024-05-15T15:29:52.447993787Z",
    "modifiedTime": "2024-05-15T15:33:26.764828611Z",
    "queued": false,
    "processedTime": "2024-05-15T15:33:21.088679723Z"
  }
}
```

Take home messages

- CDCB is investing heavily in the future of dairy industry
- Modernizing infrastructure (both hardware and software)
- Special attention to continuity and seamless integration
- Documenting all processes
- Future enhancements easier to implement
- **Work is in progress!**

*Disclaimer: Picture **not** approved by CDCB communications and outreach specialists*

THANK YOU FOR YOUR ATTENTION

ACKNOWLEDGMENTS

U.S. dairy producers
Member sectors and collaborators
USDA AGIL
CDCB staff