# CDC Data Hub Advanced Analytics Data Assets

GitHub Repositories

* [cdcent/cdh\_mlops](https://github.com/cdcent/cdh-mlops) – compute, package requirements, utilities, documentation
* [cdcent/cdh\_featurization](https://github.com/cdcent/cdh-featurization) – Medallion silver zone data engineering code and documentation
* cdcent/cdh\_adv\_analytics – Parameterized methods and data transforms

Databricks Schemas and Volumes

* Azure Group ***gp-u-EDAV-CDH-ADVANCED-ANALYTICS-DBR-ANALYST*** to control Analyst Access to all DB Assets
* [cdh\_ml](https://adb-5189663741904016.16.azuredatabricks.net/explore/data/edav_prd_cdh/cdh_ml?o=5189663741904016) – All ml data assets that are not dataset specific
  + Tables
    - concept\_set – shared concept set
    - ner entity vocabularies e.g. vocab\_biomedical\_ner\_all
    - umls\_releases – urls of umls releases for automated download
  + Volumes
    - [**metadata\_compute**](https://adb-5189663741904016.16.azuredatabricks.net/explore/data/volumes/edav_prd_cdh/cdh_ml/metadata_compute?o=5189663741904016)
      * *packages* directory with subdirectories for wheels and jars
      * requirements\*.txt referencing packages for cluster library configuration or notebook/workflow use
    - [**metatdata\_data**](https://adb-5189663741904016.16.azuredatabricks.net/explore/data/volumes/edav_prd_cdh/cdh_ml/metadata_data?o=5189663741904016&volumePath=%2FVolumes%2Fedav_prd_cdh%2Fcdh_ml%2Fmetadata_data%2Fathena%2F)
      * *cache* directories for complex ML/AI packages
      * directories for support downloads
        + UMLS
        + Athena
      * Serving directories
        + ONNX models
        + SpaCY Entity Linker UMLS data
      * Support scripts and documentation
* **cdh\_<dataset>\_ra** schemas – Dataset specific support for Retrospective Analysis (Read Only to analysts)
  + **Note** tables
  + **Fact** tables
  + **Note\_NLP** tables
  + Any other published support tables that are dataset specific
* **Workflows** – [Parameterized Workflows Tagged CDH\_AA](https://adb-5189663741904016.16.azuredatabricks.net/jobs?o=5189663741904016&tags=%22CDH_AA%22%3A%22%22) referencing cdcent repos
* **Compute**
  + [Databricks CPU ML All Purpose Cluster](https://adb-5189663741904016.16.azuredatabricks.net/compute/clusters/0630-181015-713pi206?o=5189663741904016)
  + [Databricks GPU ML All Purpose Cluster](https://adb-5189663741904016.16.azuredatabricks.net/compute/clusters/0630-181430-yfox4zns?o=5189663741904016)
  + Databricks Cluster Policies
    - SparkNLP (GPU)
    - Huggingface (GPU)
    - PYMC (CPU)
    - Generic CPU ML