Aggregat - Roadmap

High priority

1. **Migrate from DataOps PoC ("TEST") to DEV environments**
    - DEV Data Bank
        - Already exists
    - DataPrep "Weather Data Extraction" (Grib)
        - COPY old data ("Solvay Energy Aggregation - Parsed Grib") to DEV Data Bank (DataBank Dev - Weather Forecast) to keep history
            - check README for SQL procedure
    - DataPrep "SES Aggregat DataPrep"
        - Migrate to DSS "DEV" (Design & CI Automation) - import/export
            - recreate connections if needed
        - Update Gitlab CI variable ("ci infrastructure key")
    - DataApp "SES Aggregat"
        - Migrate to DSS "DEV" (Design & CI Automation) - import/export
            - recreate connections if needed
        - Update Gitlab CI variable ("ci infrastructure key")
    - DataPrep Pipeline Scheduling
        - Create DEV environment on GitLab project based on dataprep_pipeline_test_env
        - Configure it to use DEV projects (DataPrep "Weather Data Extraction" & DataPrep "SES Aggregat DataPrep")
            - CI/CD variables DATAIKU_AUTOMATION_API_KEY and DATAIKU_AUTOMATION_URL
            - project ids
        - Remove this TEST environment
    - DataApp Pipeline Scheduling
        - Create DEV environment on GitLab project based on dataapp_pipeline_test_env
        - Configure it to use DEV projects (DataApp "SES Aggregat")
            - CI/CD variables DATAIKU_AUTOMATION_API_KEY and DATAIKU_AUTOMATION_URL
            - project id
        - Remove this TEST environment
    - **Update Documentation about DEV environment**
2. **Increment Aggregat DataOps with missing evolutions**
    - Evolutions and fixes done during DataOps should be added to Aggregat new projects
3. **Integrate DataApp outputs in Energy DB**
    - Use GBQ databases as input of insert procedure
4. **Go to PROD**
    - PROD Data Bank
        - Create it
        - Review Access
    - DataPrep "Weather Data Extraction" (Grib)
        - COPY old data ("Solvay Energy Aggregation - Parsed Grib") to PROD Data Bank (DataBank Dev - Weather Forecast) to keep history
            - check README for SQL procedure
        - Change output to PROD Data Bank (prod config file)
    - DataPrep "SES Aggregat DataPrep"
        - Change "prod infrastructure key" to deploy on right environment (PROD Automation)
        - Handle PROD Automation connections
            - output on PROD Data Bank
    - DataApp "SES Aggregat"
        - Change "prod infrastructure key" to deploy on right environment (PROD Automation)
        - Handle PROD Automation connections

        - input from PROD Databank
        - output on PROD GBK Databases
    - DataPrep Pipeline Scheduling
        - Create PROD environment on GitLab
        - Configure it to use PROD projects
            - CI/CD variables DATAIKU_AUTOMATION_API_KEY and DATAIKU_AUTOMATION_URL
            - project id
    - DataApp Pipeline Scheduling
        - Create PROD environment on GitLab
        - Configure it to use PROD projects
            - CI/CD variables DATAIKU_AUTOMATION_API_KEY and DATAIKU_AUTOMATION_URL
            - project id
    - **Update Documentation about PROD environment**
5. **PoV architecture removal**
    - Remove SES Aggregat previous DSS app
    - Remove Grib & Google sources
6. **DataPrep (Weather - Grib) deployment improvement**
    - In PoC, the docker container runs within a shared Gitlab Runner.
        - Cloud Run in GCP has been considered and tested but would require a lot of refactoring to split the logic, and reassemble the extraction as a collection of micro services. Cloud run is made to deploy webapps, not long running scripts.
        - A dedicated Gitlab Runner should be assigned for this type of work. This runner could be shared between Extraction containers.
7. **DataApp (SES Aggregat) deployment improvement**
    - ML Models are automatically retrained every week on deployed automation node.
    - ML Models are not retrained on design node.
    - When a new deployment is made, the models available on the Design node will override the models previously trained on the Automation node.
    - A manual trigger of the train scenario on the target Automation node is required after a deployment to keep up to date models
    - This step should be added to the CI/CD pipeline, after the prod_deploy step.
        - Ideally, only the model needs to be retrained, is it not required to rebuild the dependent datasets if the project was already deployed.
    - Note : This behavior has been submitted to the Dataiku support through this ticket, and may change in the future:
8. **DataApp (SES Aggregat) functionnal testing performances**
    - Will apply to CI and CD
    - Predict scenario is run daily, it makes a prediction by based on the current day's weather forecast data and also based on the last 14 days' weather forecast data history. Unfortunately, accessing weather data for the last 14 days has to pull the entire history and that takes about 40 minutes! Weather forecast data is currently split in 2 datasets (current day, history). It should be splited in 3 datasets (current day, last 14 days, history). And history dataset should only be used for training models.
    - Currently the running time of the functional test add a lot of latency to deployment which could be drastically reduced

Medium priority

1. **Documentation migration to Confluence**
    - Should be migrated to a Confluence project and shared with everyone
2. **DataApp integrate more business logic**
    - Split logic and processing of insertPrevisionElec between DSS DataApp dedicated Flow Zone and rework procedure
    - Add needed tables into DataPrep Energy & Data Bank if data is not enough
3. **DataPrep (Weather - Grib) performance**
    - Each day, it takes between 1 and 2 hours to process daily ARPEG or AROME Weather forecast data.
    - Weather Data Extraction process can easly be parallelized by models (already done), files and columns. On a VM with more CPU processing can be reduce by x10 or more.
    - The Weather Data Extraction docker image can be run in kubernetes environment with autoscaling and pay per use capabilities, so parallelization over cost will faster and cheaper at least.
4. **DataPrep (Solvay Energy Databases) migration to Talend**
    - As a whole, this application is just an ETL reading from Solvay Energy databases, and writing in the Databank. That kind of work should be done with Talend.
5. **DataPrep (Weather - Grib) missing days handling**
    - If all failed during 1 day, let the script download grib files from CloudStorage on passed dates if error occurs

Low priority

1. **DataPrep (Weather - Grib) Sources paid account**

    - At this point, there is no guarantee that MeteoFrance will provide these forecast in the same format forever. The criticality of this source of data should be assessed by PO with all stakeholders. Paid subscriptions with better guarantees are available from Meteo France.
2. **DataApp (SES Aggregat) improve CI & Run health tests**
    - Post tests checks could be added, like time measurement and logs feedback
3. **RTE data (Solvay Energy Databases) update**
    - are outdated since 2019. This data source should be updated.
4. **DataPrep (Solvay Energy Databases) Incremental synchronisation => DONE 2022-07-01 during DataOps PoC project**
    - It is not possible with DSS v9.0 to incrementally append to GBQ. Once platform is migrated to v10.0, it should be refactored. See Dataiku tickets for more information.
    - Should done in Talend if `DataPrep (Solvay Energy Databases) migration to Talend` was done
5. **DataApp (SES Aggregat) AROME CI**
    - Add some AROME unit test
    - Priority could be incremented if AROME is used
6. **DataPrep (Weather - Grib) migration to Talend**
    - This application download weather forecast daily from Meteo France Open API, saves it in the Datalake, process it and saves it in the Databank. This type of work could be done within Talend if the Grib files can be processed.