Wekelijks verslag

Stage ACA group

IT factory

Kyano Trevisan

Academiejaar 2024-2025 Campus Geel,

Kleinhoefstraat 4, BE-2440 Geel

LID VAN ASSOCIATIE KU LEUVEN

THOMAS MORE

Inhoudstafel

Week 1 (24/02/2025 – 28/02/2025)

This week marked the start of my internship at ACA group, where I will be working as a data engineer on the sustainability reporting for Duvel Moortgat. The week involved onboarding, introduction to our project, a site visit to Duvel's headquarters in Puurs, and beginning technical training on the technology we will be using for the project.

Key Accomplishments:

- Completed onboarding and introduction
- Visited Duvel Moortgat's facility in Puurs for introduction and a small tour
- Started dbt (data build tool) training through Udemy
- Reviewed Duvel's logical data architecture and discussed potential approaches
- Set up development environment: dbt-core, conda, Azure VM
- Began exploring Duvel's master data to understand the available datasets
- Started Microsoft Fabric tutorial to prepare for implementation

Challenges:

- Azure License Delays: We received MacBooks which require a virtual machine setup for Power BI. The Azure license acquisition process took the entire week (from Feb 24/25 until Feb 28), which delayed our complete environment setup.
- dbt Environment Setup: The same Azure license delay affected our ability to complete the dbt environment setup, as we need to authenticate with Azure CLI.
- Data Architecture Changes: The logical data architecture we're working with isn't fully current, as several changes have been made since it was initially drafted. This creates some uncertainty about the exact data flows.

Next Steps:

- Continue with dbt training
- Get Microsoft Fabric Analytics Engineer Associate certification
- Take Microsoft PowerBI training

WEEK 2 (03/03/2025 – 07/03/2025)

During my second week at ACA, the focus shifted to technical understanding and project architecture. I had several meetings, including the kickoff with my Thomas More supervisor, and architecture discussions with the client. Most of my time this week was spent continuing my DBT training to build the necessary skills for the project.

Key Accomplishments:

- Attended an architecture meeting with Duvel stakeholders to discuss technical approaches and data sources
- Made good progress with my DBT training
- Debugged an integration notebook
- Researched up-to-date CSRD (Corporate Sustainability Reporting Directive) regulations
- Created a reference for CSRD requirements that applied to the project
- Identified key stakeholders responsible for different data domains

Challenges:

- Integration notebook issues: I spent a decent amount of time debugging the integration notebook. The root cause was eventually identified as Microsoft Fabric's aggressive caching behavior, which continued to return the same error message even after I had fixed the underlying issue (which was a human error in an input template). This caching behavior made troubleshooting more complex and time-consuming than anticipated.
- Architectural clarity: The data architecture is complex, but after meeting with some key people who are actively working with the data on our client’s side most of our questions and doubts were cleared up.
- CSRD Regulation Research: Finding the most current CSRD regulations also proved difficult, requirements seem to be constantly evolving, and documentation is spread across multiple sources.

Next Steps:

- Finish DBT training
- Continue working towards the Microsoft Fabric Analysis Engineer Certification
- Follow up with stakeholders for specific data sources:
    - Wien for solar panel data
    - An for waste management data
    - Bart for Salesforce & Microsoft Dynamics information
    - Gunter for production data
- Further identify specific HR data points needed from Workday CSRD compliance
- Explore the Engie portal export that Dries will share with the team
- Compare Engie data with Powerpulse data to determine the best answer

Week 3 (10/03/2025 – 14/03/2025)

This week marked a significant milestone as I completed my DBT training and progressed through the Microsoft Fabric course. We worked at different locations – onsite at Duvel's headquarters in Puurs on Thursday, and at ACA's office in Leuven on Friday. A productive client meeting on Thursday afternoon allowed us to discuss our progress and align on data challenges.

Key Accomplishments:

- Completed the DBT course, gaining comprehensive understanding of the data build tool environment
- Made substantial progress on the Microsoft Fabric course, learning about:
    - Data lakes architecture and implementation
    - PySpark functionality and applications
    - Delta Lake table's structure and benefits
    - Pipeline creation and management
    - Dataflow configuration and optimization
- Participated in a client meeting to discuss progress and challenges with the sustainability reporting
- Worked from multiple locations, gaining exposure to different work environments within the project scope.

Challenges:

- Data Source Integration: We identified some limitations with certain data sources that only provide partial information, creating challenges for comprehensive reporting across all sites
- Data Standardization Needs: Inconsistencies in naming conventions and data transformations will require additional effort to standardize for reliable automated reporting
- Dashboard Refinement: The current PowerBI dashboard requires restructuring to better align with strategic indicators and provide more meaningful insights
- Manual Workarounds: Some data points currently rely on manual input sheets as temporary solutions until automated queries can be properly configured

Next Steps:

- Complete the Microsoft Fabric course to gain comprehensive understanding of the platform
- Begin addressing the data standardization challenges identified in the client meeting
- Develop a strategy for consistent data naming across sources
- Create a plan to improve the current codebase by replacing hardcoded values with parameters
- Design prototype improvements for the PowerBI dashboard incorporating strategic indicators
- Investigate solutions for query issues identified during the client meeting

WEEK 4 (17/03/2025 – 21/03/2025)

This week marked a transition from training to hands-on implementation work. I completed the Microsoft Fabric course and began applying my knowledge to solve real project challenges. My focus shifted to building and improving data pipelines, enhancing error handling, and implementing security best practices. I also participated in meetings with key stakeholders to address specific data issues.

Key Accomplishments

- Completed the Microsoft Fabric course, gaining comprehensive understanding of the platform
- Created a transformation pipeline to streamline data processing
- Implemented security improvements by moving client credentials to KeyVault instead of hardcoding them
- Developed a proof of concept for pipeline failure notifications in Microsoft Teams
- Created documentation for pipeline failure options for the client to choose from
- Resolved NaN issues by modifying Excel import settings and improving numeric conversion logic
- Debugged several components including dbt code and sitedata errors
- Participated in meetings with Dries and Gunter to address specific data challenges
- Added a new data source for Scope 1 & 2 data, filling a previous data gap

Challenges

- Data Type Conversion: Encountered issues with NaN values in data imports that required explicit null value handling and modifications to Excel import settings.
- Pipeline Reliability: Needed to develop better error handling and notification systems to ensure data pipelines run reliably.
- Security Concerns: Identified hardcoded credentials in configuration files that needed to be moved to a more secure solution.
- Code Debugging: Several components required debugging, including dbt code and transformation notebooks.
- There were sheets in which people had skipped several rows, this resulted in warnings, I documented which sheets produced errors

Next Steps

- Implement the Teams notification system for pipeline failures in the production environment
- Continue improving error handling and data validation in the notebooks
- Expand the data transformation pipeline with additional data sources

Week 5 (24/03/2025 – 28/03/2025)

This week was heavily focused on data modeling and analysis, with significant progress made on the transformation pipeline and indicator tables. I participated in several meetings, including a steerco (Steering Committee meeting), and discussions about API integrations. The week involved solving several technical challenges, as well as conducting research on specific data points.

Key Accomplishments

- Advanced the data modeling through ERD (Entity Relationship Diagram) design and identified potential improvements
- Improved the transformation pipeline
- Fixed integration code for pillar tables to improve data reliability
- Fixed filtering issues for BaseUnitOfMeasure and InvoiceTypeItems in production volumes
- Created and refined indicator tables for automated reporting for certain "pillars"
- Completed the integration of Scope 1 & 2 (emissions) data
- Researched site data issues to determine root causes, and fixed them
- Participated in the first Sustainathon steerco
- Team sync to align on current progress and challenges
- Powerpulse meeting to review Fluvius API integration for sprints 3-4
- Meeting about Workday data, Datawarehouse, and Failure Notifications

Challenges

- Authentication Issues: Encountered problems with semantic model authentication, this was not a major issue,
- Site Data Problems: Identified and researched issues with site (V01, Bernard, Mont Blanc, t'Ij, CN) data that required special handling

Next Steps

- Address authentication issues in the semantic model
- Continue ERD modeling to support improved data architecture
- Simplify and clean up the current data model, adhering to the medallion structure principles
- Implement development and production environments on Microsoft Fabric, as large changes in the data structure will break the dashboard

WEEK 6 (31/03/2025 – 04/04/2025)

This week was focused on solving specific data quality issues and implementing architectural improvements for the sustainability reporting platform. I worked on several technical solutions, including a reference-based approach for non-alcoholic beer identification and participated in an onsite meeting with the client where we established plans for test and production environments. Significant progress was made in the planning of the improved data model.

Key Accomplishments

- Resolved cloud connection authentication issues that were affecting semantic model updating
- Updated input sheets with improved data validation
- Participated in architecture planning for test and production environment setup, since when I implement the improved data model, visualizations in the current dashboard will have to be updated
- Modified Purchase and Sales queries to handle specific data requirements, including:
    - Weight calculation for products with zero weight data
    - Specific data points as required by business rules
    - Customer shipping address logic

Challenges

- Data Quality Issues: Many beer products had incorrect alcohol percentage values in source systems, requiring a custom reference-based solution, we are still working this out
- Cloud Connection Issues: Discovered authentication problems with semantic model connections due to conflicts between personal Microsoft accounts and Service Principal configurations
- External Data Sources: We are currently waiting on our client to provide us with API access to Energy and Solar Panel data, so we can implement these into our dashboard

Next Steps

- Implement the test environment setup according to the architecture plan
- Add a new site to the input sheets
- Continue refining the purchase and sales queries
- Automate certain transformations that are currently being performed manually by the client
- Create Production workspace (for Duvel’s Operational data)

WEEK 7 (21/04/2025 – 25/04/2025)

This week focused on debugging and enhancing the data transformation pipeline with improvements to the location code handling and emission factor calculations. I also began implementing the test/production environment setup to improve development workflow. Note that Monday (April 21) was Easter Monday, so this was a four-day work week.

Key Accomplishments

- Fixed multiple issues in the transformation pipeline that were causing failures
- Implemented a more dynamic approach for location code assignment across data sheets
- Added NULL handling functionality to improve data quality and pipeline reliability
- Modified emission factor equivalent data types to support mathematical calculations
- Began implementing the test/production environment setup based on Microsoft Fabric best practices
- Attended a meeting with ClimateCamp to understand their data transformation processes
- Implemented empty row handling as sometimes rows were left empty, or partially empty in the input sheets

Challenges

- Excel Import Edge Cases: Encountered and resolved an unusual issue where hidden "Unnamed" columns were being imported from a specific Excel sheet, causing the location code assignment to fail
- NULL Value Handling: Spent significant time debugging why some rows weren't receiving location codes properly
- Semantic Model Refresh Issues: Had to troubleshoot and fix naming inconsistencies that were preventing the semantic model from refreshing

Next Steps

- Complete implementation of the test/production environment setup
- Apply transformations learned from ClimateCamp meeting
- Add additional data validation checks to prevent future pipeline failures
- Continue refining the DBT code for better performance and reliability
- Implement additional sanity checks for data quality

Week 8 (28/04/2025 – 02/05/2025)

This week was highly productive with significant focus on setting up a testing and production environment alongside ongoing pipeline improvements. I attended several strategic meetings, including the second steerco and a session with HR to discuss sustainability indicators. A substantial portion of the week was dedicated to implementing the test/production environment setup, including parameterizing pipelines and resources.

Key Accomplishments

- Participated in the second steerco meeting for the project
- Contributed to architecture discussions about environment setup for multiple data domains
- Implemented the test/production environment setup with parameterized resources
    - Parameterized pipelines to accept lakehouse and warehouse IDs as variables
    - Created configuration for single-click deployment between environments

Challenges

- Permission Limitations: Had to adapt implementation plan due to restricted permissions for workspace creation
- Parameterization Complexity: Ensuring all components properly use parameters required significant refactoring
- Architecture Decision Making: Navigating the trade-offs between different architectural approaches for environment organization

Next Steps

- Apply lessons learned from ClimateCamp meeting to optimize data transformations
- Continue refining the DBT code for better performance and reliability
- Implement additional sanity checks for data quality

WEEK 9 (05/05/2025 – 09/05/2025)

This week focused on significant code quality improvements through dynamic programming techniques, database connectivity, and ongoing ClimateCamp transformations. I implemented several advanced solutions that enhance the maintainability and scalability of the data pipeline, particularly through the use of Jinja templating in DBT to create more dynamic and reusable code structures. I also helped resolve database connection issues and improved the currency conversion process to support year-specific rates.

Key Accomplishments

- Implemented dynamic location source prefix handling using Jinja lists in DBT
- Created a year-based currency conversion system that automatically uses the correct historical conversion rates
- Connected to the client's new water database through gateway configuration
- Removed hardcoded values throughout the codebase, improving maintainability
- Cleaned up integration notebooks for better readability and exception handling
- Resolved Microsoft SQL Integrated Security connection issues for the client
- Advanced ClimateCamp transformations and address handling

Challenges

- Database Connectivity: Navigating the gateway setup process required coordination with the client to establish proper permissions and configuration
- Dynamic Programming in SQL: Implementing Jinja templating effectively required careful consideration of SQL generation patterns
- Historical Conversion Rates: Creating a system that automatically uses the correct year's conversion rates while maintaining extensibility for future years
- Authentication Issues: Resolving Microsoft SQL Integrated Security problems required identifying the difference between Windows and Microsoft account authentication

Next Steps

- Continue refining ClimateCamp transformations
- Document the new dynamic source handling approach for future maintainers
- Create a process document for yearly updates to conversion rates
- Implement additional validation checks for the new water data source

WEEK 10 (12/05/2025 – 16/05/2025)

This week marked the completion of several major components of the sustainability reporting platform. I finalized the ClimateCamp data transformations that we've been working on for several weeks, ensuring all data flows correctly into their system for environmental impact calculations. Additionally, I completed a comprehensive refactoring of the transformation codebase to implement a proper medallion architecture, significantly improving data quality and structure.

Key Accomplishments

- ClimateCamp Integration Completion: Finalized all data transformations required for ClimateCamp integration, ensuring proper formatting and structure for environmental calculations
- Major Codebase Refactoring: Completely restructured the transformation pipeline to implement a proper medallion architecture (Bronze, Silver, Gold layers)
- Data Quality Improvements: Ensured all data flows through the correct validation and transformation steps with improved accuracy
- Documentation and Code Cleanup: Created comprehensive documentation and cleaned up legacy code for better maintainability

Challenges

- Architecture Migration: Migrating existing data flows to the new medallion structure required careful planning (This is what we need the Test/Prod environments for)
- Data Validation: Ensuring data accuracy across all transformation layers while maintaining performance
- ClimateCamp Requirements: Meeting specific formatting and calculation requirements for environmental data processing

Next Steps

- Continue documentation efforts for all transformation processes

WEEK 11 (19/05/2025 – 23/05/2025)

This week was dedicated to comprehensive documentation, final code optimization, and preparing the sustainability reporting platform for production use. Building on the medallion architecture refactoring completed last week, I focused on creating detailed documentation for all processes and conducting thorough testing to ensure data accuracy and system reliability.

Key Accomplishments

- Comprehensive Documentation: Created detailed documentation covering all aspects of the data transformation pipeline, including architecture diagrams, process flows, and maintenance procedures
- Code Optimization and Cleanup: Finalized code cleanup efforts, removing deprecated functions and optimizing performance across all layers of the medallion architecture
- Data Validation Testing: Conducted extensive testing to verify data accuracy and consistency throughout the transformation pipeline
- Process Documentation: Documented operational procedures for ongoing maintenance and troubleshooting
- Knowledge Transfer Preparation: Organized all documentation and code for effective handover

Challenges

- Documentation Scope: Ensuring documentation was comprehensive enough for future maintainers while remaining accessible and practical
- Knowledge Capture: Documenting not just what the code does, but why certain decisions were made for future reference

Next Steps

- Conduct final end-to-end testing of the complete system
- Prepare presentation materials for project completion
- Finalize any remaining minor optimizations
- Prepare for knowledge transfer sessions with the client team
- Complete final validation with stakeholders

WEEK 12 (26/05/2025 – 28/05/2025)

This final week of my internship focused on completing the knowledge transfer process, conducting final system validation, and preparing comprehensive handover materials. I worked closely with the client team to ensure a smooth transition and completed final testing of the entire sustainability reporting platform. The week concluded with successful delivery of a fully functional, documented, and production-ready system.

Key Accomplishments

- Knowledge Transfer Sessions: Conducted comprehensive handover sessions with the client team, covering all aspects of the sustainability reporting platform
- Handover Documentation: Finalized all documentation packages, including technical specifications, user guides, and maintenance procedures
- Project Completion: Successfully delivered a complete sustainability reporting solution meeting all CSRD compliance requirements
- Stakeholder Presentations: Presented final results to project stakeholders and received approval for production deployment.