Sebastian Daberdaku, Ph.D. | Data Engineering Tech Lead @ Cardo AI | Data Platform

---

portfolio **publications** **LinkedIn** **Medium**

Summary

With over 5 years of experience as a data engineer and a PhD in Information Engineering, I specialize in crafting efficient, scalable, and resilient data pipelines. Proficient in technologies like Spark, Airflow, Python, DataBricks, TrinoDB, Kafka, Debezium, and AWS, I've worked across diverse domains including bioinformatics, clinical informatics, and finance. My research and data science background ensure a rigorous problem-solving approach. I am committed to ongoing learning and staying updated with the latest advancements in data engineering to drive impactful insights.

Experience

Aug 2023 – present Data Engineering Tech Lead @ Cardo AI

- Implemented a **centralized multi-cluster Kubernetes (EKS) monitoring solution** with the LGTM stack (Loki, Grafana, Tempo, and Mimir).
- Designed and implemented a **centralized AWS logging solution** using CloudWatch, Subscription Filters, Kinesis Firehose Data Stream, and S3 for log streaming, ensuring security, immutability, and cost-efficient storage across multiple AWS accounts.
- Implemented a **fully automated Change Data Capture (CDC) with schema evolution support of ~2k RDS Postgres tables to Delta Lake on AWS** with Kafka, Debezium, Confluent Schema Registry, Spark Connect and Airflow on EKS, with fully automated monitoring comprised of Grafana Dashboards, Prometheus Rules and Alertmanager.
- Designed and implemented a multi-AWS-account CI/CD infrastructure management pipeline with GitHub Actions and AWS CodeBuild.
- Developed a **multi-account strategy for Cardo AI's AWS environments** according to the AWS Well-Architected Framework, for optimal data segregation, cost accounting, security and scalability. A centralized **Networking** account provides secure and private connectivity between the centralized services and the customer workloads while allowing fine-grained access controls with security groups and NACLs. A dedicated **Infrastructure Management** account is responsible for deploying and managing the cloud infrastructure while enforcing strict security rules. A **Common Workloads** account serves the shared services to multiple and independent **Customer Workload** accounts, ensuring cost-effective usage of resources, and data segregation. Finally, the **Log Archive** and **Security Tooling** accounts provide the means to securely and immutably store logs, metrics, traces, and to centrally operate security services.
- Implemented and managed Cardo AI's Data Science cloud infrastructure (AWS) and tooling on EKS with IaC (Terraform) and GitOps (ArgoCD), including automated **KubeFlow** deployment with **deployKF** and dedicated Kubernetes Job, **MLflow** deployment, and EKS node provisioning with

Karpenter NodePools for correct workload segregation (spot vs. on-demand, AMD vs. ARM, GPU nodes).
- Developed a **Helm Chart for deploying cost-effective and stable TrinoDB clusters on EKS with Karpenter**, using EC2 spot instances for the Trino workers, automatic query retry policy, spill-to-disk and Alluxio cache support on NVMe.
- Refactored the synchronization process of Cardo AI's structured finance suite (**Equalizer**) using Airflow, Spark Connect and Delta Lake, enhancing parallelism and scalability.
- Implemented fully-automated deployments of Apache Airflow and Databricks environments with Terraform modules and ArgoCD for various teams in the company, guaranteeing data separation and detailed cost allocation.
- Developed a **Helm Chart for Spark Connect server and Spark Thrift server**, which can be used concurrently by multiple Apache Spark applications, effectively optimising infrastructure costs and improving the overall developer experience.
- Ensured that the Data Engineering team followed best practices and maintained high-quality code and data pipelines.

Sep 2022 – Jul 2023 Senior Data Engineer @ Cardo AI

- Refactored the whole Cardo AI Data Engineering infrastructure on the AWS cloud. Implemented a multi-environment setup with automated CI/CD and IaC principles.
- Developed several Terraform modules for providing complex services to our stakeholders. **Tech stack**: AWS, K8s (EKS, Karpenter, Helm Chart, ArgoCD, Vault) TerraformHub Actions.
- Developed an **on-demand big data processing** solution with Airflow – Spark – Delta Lake – K8s – Trino, released as an **Apache Airflow provider package**. **Tech stack**: K8s, Apache Airflow, Apache Spark, Delta Lake, AWS Glue (as meta-store), and Karpenter.
- Developed and maintained several **ETL pipelines** with Apache Airflow and Databricks. **Tech stack**: Python, Apache Airflow, Databricks, Apache Spark (PySpark), Delta Lake.

May 2020 – Aug 2022 Senior Data Scientist & Engineer @ Sorint.Tek

- **AWS Data Architect** for the Data Engineering team at InfoCert S.p.A where I developed and improved automated data ingestion processes. **Tech stack**: AWS Infrastructure (CloudFormation), CI/CD, Lambda, RedShift, Kinesis, Glue, RDS, S3, Python, fluentbit.
- **Baggage Carousel Assignment** at the Milan Bergamo Airport. Developed an automatic scheduling algorithm for the optimal assignment of flights to baggage belts in the baggage reclaim area of the airport, Client: SACBO S.p.A. **Tech stack**: Python (OR-Tools, SQLAlchemy), Oracle, Docker.
- **SNIPE** (Sensor Network for Intelligent Predictive Enterprise). Developed an AI enabled IoT system for machinery revamping in foundries to monitor performance and enable predictive maintenance based on **MangroviaIoT**, Client: FAE Technology S.p.A. **Tech stack**: Python (scikit-learn), TimescaleDB.
- **ELK SEA**. Developed an analytics platform for anomaly detection and monitoring on the WiFi networks of the Linate and Malpensa airports (Milan, IT), Client: SEA SpA. **Tech stack**: Elasticsearch, Kibana, Logstash, RabbitMQ, Cisco Prime, Python (requests), bash.
- **ELK ML OpenShift**. Developed machine learning models for microservice anomaly detection and root-cause analysis on an OpenShift cluster, Client: Intesa Sanpaolo Group Services S.C.p.A. **Tech stack**: Elasticsearch, Kibana.
- **2Vita-B**. Developed deep-learning models for stress and affect detection from wearable sensor data, Client: Atlantica Digital S.p.A. **Tech stack**: Python, Keras Tensorflow (CNN), scikit-learn,

AWS SageMaker.

- **Tech Lead** for **Safe LTA**, a cloud-based long-term repository of digital documents, Client: InfoCert S.p.A. **Tech stack**: Java (Spring Boot), Docker, ElasticSearch, Amazon S3, Kong API Gateway, Keycloak, RabbitMQ, SQS, Elastic Kubernetes Service, AWS Architecture (CloudFormation).

Jan 2016 – Apr 2020 Postdoctoral Researcher @ University of Padua

- **Data Mining from environmental, genetic and clinical variables in ALS**, Dept. of Information Engineering.
  - ALS prognosis prediction with machine learning and data mining methods. **Tech stack**: R (bnstruct), R-Shiny, Dynamic Bayesian Networks.
  - Missing data imputation for static and dynamic mixed-type clinical data. **Tech stack**: R (e1071, caret, infotheo).
  - Metagenomics diagnosis for Inflammatory Bowel Disease. **Tech stack**: R, SVM, Randomised Logistic Regression.
- **Evolutionary-based approach for predicting protein interaction sites and residue mutation impact**, Dept. of Comparative Biomedicine and Food Science.
  - Single-point-mutation effect prediction on metabolic networks with flux balance analysis and machine learning methods. **Tech stack**: Python.
- **Models and algorithms for protein–protein docking**, Dept. of Information Engineering.
  - Protein interface prediction with machine learning methods. **Tech stack**: C++, Python (scikit-learn).
  - Protein surface representation for the docking problem. **Tech stack**: C++ (boost, MPI, Open-MP).
  - Parallel computation of molecular surfaces with MPI/OpenMP. **Tech stack**: C++ (boost, MPI, Open-MP).
  - Protein pocket and cavity identification. **Tech stack**: C++, Python.

Jan 2013 – Apr 2020 Lecturer @ University of Padua

- *Computer Engineering Laboratory* course (C programming), B.Sc. in Information Engineering.
- *Foundation of Mathematical Analysis and Probability* course (tutoring), B.Sc. in Biomedical Engineering.
- *Probability and Statistics* course (tutoring), B.Sc. in Mathematics.
- *Embedded Systems Programming* course (tutoring), M.Sc. in Computer Engineering.

Professional Qualifications

- *22/07/2024* **Astronomer Certification DAG Authoring for Apache Airflow**
- *11/01/2024* **Astronomer Certification for Apache Airflow Fundamentals**
- *29/11/2021* **AWS Certified Solutions Architect – Associate**
- *19/07/2021* **Elastic Certified Analyst**
- *21/01/2013* License to practice the profession of *Information Engineer*, University of Padua

Education

2013 – 2016 Doctor of Philosophy

- Information Engineering, University of Padua

- Thesis: *Protein contour modelling and computation for complementarity detection and docking*.

2010 – 2012 Master's degree

- Computer Engineering, University of Padua
- 110/110 *cum laude*
- Thesis: *DHT-based task allocation in Volunteer Computing*.

2007 – 2010 Bachelor's degree

- Computer Engineering, University of Padua
- 110/110 *cum laude*
- Thesis: *PariMulo: Credits*.

Awards and Achievements

- *October 2024* Trino Champion nomination for the **"several awesome contributions to the Trino Helm Chart"**.
- *13/07/2020* Winner (Team GiGi) of the **"Metagenomics Diagnosis for IBD Challenge" (MEDIC)**.
- *21/11/2018* Winner of the PeerJ Award for the **"Best Contribution by an Early Career Researcher" @ BBCC2018** for the paper "Identification of protein pockets and cavities by Euclidean Distance Transform".
- *2013 – 2016* Full Ph.D. scholarship, University of Padua.

Personal Data

I hereby authorize the use of my personal data in accordance with GDPR 2016/679 (General Data Protection Regulation).

Other formats

`resume.html` `resume.pdf` `resume.docx` `resume.rtf`

October 29th, 2024