(+34) 626 364 652
dsaiztc@gmail.com
Medina de Pomar, Spain

DANIEL SAIZ

English / Spanish

dsaiztc.com
github.com/dsaiztc
linkedin.com/in/dsaiztc

Work Experience

Shopify (Spain, 100% Remote) 06.2021 - 08.2024 (3y, 3m)

Shopify is a leading global commerce company, providing trusted tools to start, grow, market, and manage a retail business of any size. Shopify makes commerce better for everyone with a platform and services that are engineered for reliability, while delivering a better shopping experience for consumers everywhere. Shopify powers over 1.7 million businesses in more than 175 countries and is trusted by brands such as Allbirds, Gymshark, PepsiCo, Staples, and many more.

Senior Software Engineer @ Streaming Data Infrastructure 06.2023 - 2024.08 (1y 3m)

Transitioned to the *Streaming Foundations Team* (6 ICs), focused on maintaining and evolving the **infrastructure** that powers all things *streaming* within *Shopify*: ***Apache Kafka*** and ***Apache Flink***. All deployed on ***Kubernetes*** via *GKE*.

For *Kafka*, the team managed multiple clusters, including some multi-region. On top of that, we maintained & evolved several services on top of *Kafka* for e.g.: *Topic Management* (creation, ownership, auto-scaling), *zonal*-based rollout of *brokers*, mirroring data. For 2023's BFCM (*Black Friday/Cyber Monday*), it served 29 million msg/sec at peak.

For *Flink*, the team managed two *regional Kubernetes* clusters, running a few hundred *Flink jobs* of different services and Tiers.

- The team owned the infra, so I was part of a **24/7 *on-call*** rotation for both *Kafka* (Tier 1, 99.9% availability) and *Flink* (Tier 2, 99.95% availability). This included attending *pages* (either alert-based or from other teams), housekeeping (e.g. right-sizing of clusters), client support and other maintenance tasks (e.g. review alerts & SLOs, keep an eye on costs).

- Other duties shared by the team in which I also participated were capacity & resilience planning (e.g. company-wide scale tests, run *Gamedays*) in preparation for BFCM.

- I was an early contributor to *Materialize Commerce* initiative: use *CDC*-derived *Kafka* messages, processed by *Flink* jobs that generate *Commerce Events* used to distribute state over different systems (i.e. replicate *SoT* data in real time). Participated on the *Flink*-infra side to prototype a *Flink* job that'd read from the *MySQL binlog* (permissions, networking, etc.)

- **Led the project** that migrated non-prod *Kafka* clusters to ***KRaft mode*** (& removed *ZooKeeper* service). Two in *staging*, one of them being multi-region; and two in *sandbox*.

- Participated in the project that rolled out the infrastructure to support running *Flink jobs* managed by the ***Flink Kubernetes Operator*** with *Tier* 2 requirements. Previous *Flink* infra was based on bare *Kubernetes* resources (i.e. *Deployments* and *StatefulSets*). This allowed to use *Operator*-exclusive features like autoscaler while improving DevX when deploying/managing a *Flink job*, ensuring previously available tooling (like automatic cluster failovers) were supported.

Senior Software Engineer @ Streaming Data Processing 06.2021 - 05.2023 (2y)

Joined the *Streaming Capabilities Team* (6 ICs) with the goal of building a platform on top of ***Apache Flink*** that'd help any developer at *Shopify* develop/deploy/monitor **streaming (a.k.a. real-time) data processing** jobs without hassle.

- Before consolidating into *Flink*, the team owned 4 completely different streaming-based services that were diligently deprecated. **I led the transition of one of those services**.

- We worked iteratively, based on user's feedback (e.g. 2021's BFCM Live Globe). Starting with templatized *Kubernetes* templates and a mono-repo. At the start of 2022 **I led the project to integrate it with the rest of *Shopify*'s dev tooling** (cli tools, observability, admin, etc.).

- Worked on making it a ***Tier 2*** service (99.95% availability, tighter security constraints) for BFCM 2022, which allowed teams to integrate *Flink jobs* in their *Tier 2* services.

- We also developed a thin library to provide support for in-house connectors (*Kafka*, *GCS, Bigtable*, etc.), which was based on ***Scala***, using the ***DataStream API***. I worked mainly on the infra side-of-things, but I also contributed to the library with e.g. tooling for easy latency measurements.

- We owned the infrastructure (i.e. ***Kubernetes*** clusters), the *Scala* library and all the associated templating (e.g. *Kubernetes* YAML files). We also maintained extensive documentation about the usage of the platform.

- At the end, we had two clusters in different regions, running a few hundred jobs from different teams/services and tiers. Provisioning more than 20k cores and 100TB of memory (& storage) in total. Some of the jobs handling TBs of state.

- I was part of the **24/7 *on-call*** rotation (alert-based monitoring, client support).

Spotahome (Madrid, Spain; 100% remote since March 2020) 09.2017 - 06.2021 (3y, 10m)

Spotahome is a 100% online (no visits required) rental platform, focused on mid/long term accommodation. Currently providing service in 9 countries (14 cities) with more than 11k landlords and 60k listings. More than 80k bookings generating 300MM GMV since started in 2014. Private series B company that has raised a total of $72 million, backed by Kleiner Perkins, Passion Capital and Seaya Ventures as lead investors.

Data Lead 11.2019 - 06.2021 (1y, 8m)

Transitioned to a more strategic role, leading the data architecture from the ground-up with the goal of allowing for more and better data products, including **Machine Learning** integrations within the product.

Continued the work on the company-wide data strategy, this time including core data infrastructure. Systems were failing too often, usually with manual intervention required, and stakeholders weren't happy. Also, new developments were too time-consuming and very restricted in scope.

Streamlined to a data infrastructure based on **CDC** (*Change Data Capture*) and **BigQuery**, reducing data *time-to-use* (from data creation to data available for decision making) by orders of magnitude, enabling bigger and wider data-based products. Change of data pipelines *paradigm* from ETL to ELT. Allowed to build DWH (on top of it) from scratch in ~3months. Reduced complexity/maintainability (previously was based on event-processing service built with Scala/Akka, Kafka, Kubernetes, Redshift), failures (everyday, with action required; to non-existent) and price while improving usability and extensibility.

This also allowed the deployment of Machine Learning models like the search ranking algorithm. Critical to have a well-thought data architecture to build on top of. **Python**-based models (sklearn), training scheduled on Airflow. *Raw* data stored on BigQuery, modelled into consumable entities and features with **dbt** (batch).

Continued working on the Machine Learning stack, setting up a **Flask**-based **API** on top of a **MySQL**-based **Feature Store** for online predictions, replicated from the offline, historic Feature Store on BigQuery. Still rough on the edges, but already integrated with the website in PROD, offering an API for recommended alternative listings.

Data Analytics & Engineering Manager — 05.2019 - 10.2019 (6m)

Stepped in the role after the Data Engineering Lead left the company.

- Led team of 5 BI Developers (Reporting, ETLs & DWH) and 3 Data Engineers (Data infra. + complex ETLs; Spark, Kafka), closely collaborating with Data Science.

- Define, prioritize & coordinate quarterly roadmaps and day-to-day task planning, including technology assessment.

- Define company-wide, mid-long term, Data Strategy, from KPIs definition to tech-stack (Data Lake, Data Warehouse, ETLs vs ELTs, Machine Learning); including Data Governance proposal with the goal of laying up the guidelines for any future developments what shall guarantee a minimum data quality.

- Explore different Data Architectures with a combination of S3/Redshift, with different data sourcing methods for doing CDC (*Change Data Capture*): AWS DMS, Kafka Connect. Explore different tech stacks (Python/Airflow, Redshift, BigQuery, SAAS Fivetran/Stitch).

- Work on a cross-functional team to build a *clickstream server-side* event tracking system (as alternative to *Google Analytics client-side* tracker) to analyze cross-device behavior and build internal A/B testing platform.

- Hiring manager for (also) other data profiles: *Data Engineer*, *Data Analyst*, *Data Scientist*. Conduct technical interviews and develop test-cases.

Data Analytics Manager — 05.2018 - 04.2019 (1y)

As the company grew, I transitioned to managing a newly-created team responsible for building company-wide decision making systems (official reporting, self-service data consumption). Worked as both people manager (1-1s, feedback, performance reviews) and technical lead. The goal was to re-create both Data Warehouse and reports from the ground-up, building the team along the way. Specially challenging as the core data of the company was schema-less (non-relational), sometimes distributed between multiple databases.

- Design and build Data Warehouse based on a myriad of different data sources (Domain Events, relational tables, MongoDB, APIs, etc.) on **Amazon Redshift**; including infrastructure provisioning, CI config., systems monitoring, etc. on AWS/k8s

- Design and build company-wide reporting on **Power BI** while also offering a self-service data platform on **Metabase**.

- Build and lead a team of up to 7 people with different backgrounds.

    - Assess & define the recruitment process, including technical interviews and technical tests.
    - Organize and prepare training sessions for Python-oriented ETL development.

- Define, prioritize & coordinate quarterly roadmaps and day-to-day task planning, including technology assessment.

- Define company-wide data strategy in terms of event standardization & domain language.

- Closely collaborate with Data Science: support on technology, data understanding, KPI definition, A/B test, Machine Learning initiatives

Data Analytics Engineer — 09.2017 - 04.2018 (8m)

Essentially split my time between collaborating with Product for analytical support and building the data infrastructure that allowed such analysis and more (DWH & reporting). Worked in a team of 5, reporting directly to the Head of Data.

- Exploratory data analysis with **Python**, driven by either PMs or own initiatives to improve business. **A/B Testing** experimentation support (experiment planning and posterior analysis). Machine Learning research.

- ETL development with Python and Pentaho/Kettle on top of a **PostgreSQL**-based Data Warehouse.

- Assessment and posterior deployment of **Apache Airflow** for batch processing (ETLs, scheduled reports) on **Kubernetes**.

    - Own initiative that I previously tested, proposed and led in order to take advantage of the upcoming (at the time) Kubernetes integration in v1.10.
    - This allowed us to better allocate infrastructure resources according to real consume, launching each individual ETL on its own POD and thus being able to dimension based on individual necessities.
    - Development of custom KuebernetesOperator for setup different environments automagically.

- Development of internal libraries for posterior ETL building with Python: data connections to different sources, API management, logging, etc.

- Assessment and posterior deployment of a Business Intelligence tool, including POC deployment of Tableau and Power BI.

- Report building with **Power BI** for company-wide consume, which included managing stakeholders requirements, help in KPI definition and ensure the quality of the data.

Skillset: Python, SQL, git, Docker, Kubernetes, Power BI, Apache Airflow, Apache Spark, Redshift, PostgreSQL, MongoDB, Google Analytics

CST International (Dublin, Ireland) 08.2015 - 07.2017 (2y)

CST International is a specialist market research agency focusing on employee engagement, employer branding, brand research and customer feedback.

Software Engineer & Data Analyst

As part of a small technical team (2.5 people), I worked on a myriad of things, ranging from deep analysis of employee engagement feedback to the full design and implementation of our purpose-built reporting system.

- Data cleaning and analysis (EDA, Hypothesis testing, visualizations and some basic ML algorithm for fraud detection). Data retrieving & wrangling mainly with **python** (requests, BeautifulSoup, pandas). Data analysis with **R** (tidyr, dplyr, ggplot, lm). Some experience with NLTK and sklearn for comment classification.

- Design and development of an *event-driven* data processing system in AWS with **python**: collects different sources of data (**S3**, **API Gateway**), process and performs different actions based on the content. All is based on events and queues (**DynamoDB**, **DynamoDB Streams**) that are automatically processed by different **AWS Lambda** functions. This was the result of a migration/unification of many different client-based systems (different scripts, jobs and DBs).

- Full Stack development for enhance/renew a reporting system running on **Web2py**. Using **Bootstrap** 3 for building the interface, **jQuery** (with a few plugins) for DOM manipulation and AJAX requests as well as visualization libraries like **Google Visualization**, **D3.js** or **C3.js**. Use of **Reportlab** for building PDFs. Running on top of DWH (Aurora/MySQL).

- Development of client-facing surveys with HTML/Bootstrap/jQuery built on top of S3 and CloudFront CDN, with backend running on AWS Lambda, tested end-to-end with Selenium and BrowserStack (as had to work with quite legacy browsers).

- Setup of **AWS Elasticsearch Service** integrated (proxied) within the Web2py reporting system using Searchkit UI Components. Before going for this UI kit I built a prototype from scratch with **Vue.js**.

Skillset: python, py.test, SQL, git, Web2py, Statistics, Selenium, HTML, CSS, JavaScript, jQuery, Bootstrap, R, Unix, AWS (S3, Lambda, DynamoDB, API Gateway, VPC, RDS, CloudWatch, CloudFront), Microsoft Power BI

UC3M - Charles III University of Madrid (Madrid, Spain) 12.2013 - 06.2014 (7m)

Software Engineer (Research Internship)

Led the design and development of a native **Android** app for *Bultaco Motors* (electric motorcycle manufacturer) through an R&D project with *Carlos III University of Madrid*.

GPS navigation app (based on *OsmAnd*) with real-time collection & display of telemetry data (Bluetooth). Full development of a **Machine Learning** system (ANN/MLP) in Java aiming to predict the viability of unknown routes based on historic data. Information retrieval from SOAP (IIS Server).

Skillset: Java, git, SQL, Machine Learning

Education

UC3M - Charles III University of Madrid (Madrid, Spain)

Telecommunications Engineering (BSc + MSc) 09.2008 - 09.2014

Top 15%. Main Coursework: Computer Science, Digital Signal Processing, Electronics, Communication Systems, Networking, Statistics.

Final Year Project Distinction with Honors: *Range Estimation System for Electric Vehicles with Artificial Neural Networks*, available here (in Spanish): http://dsaiztc.com/PFC.pdf