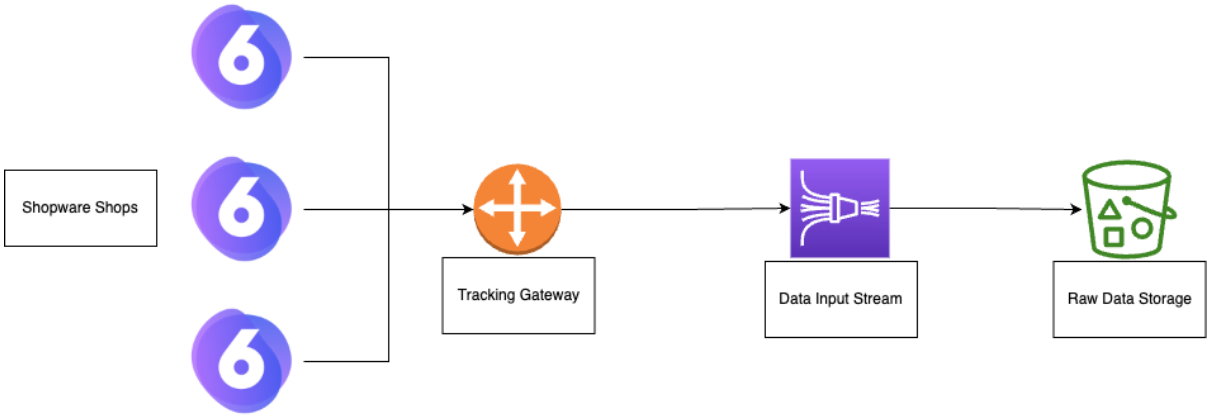
(Senior) Data Engineer Task

Setup AWS Data Infrastructure

As a Data Engineer at shopware with a focus on AWS architecture and big data, you will be responsible for designing, developing, and managing data pipelines and architectures to efficiently process, store, and analyze large volumes of structured and unstructured data. Your expertise in AWS services and big data technologies will contribute to developing AI-driven features for our product and deriving data-driven insights.

The following tasks are representative of the tasks you will find at the Data & AI department:

- **Task 1: Architecture and services on AWS**
  - Scenario: We are tracking data regarding onsite user behavior in thousands of online shops using our product. As a Data Engineer, you must decide which services to use and how those services fit into our target architecture. We have already built and implemented a so-called tracking gateway connected with an AWS Kinesis Firehose service to persist incoming data stream from those online shops into a S3 Bucket.



  - 1. Question: What are the potential challenges and risks of using this architecture and services in the given scenario? Hint: Traffic will be very high, and it will increase exponentially.

  - 2. Question: What architecture or services do you recommend for further processing the data to check data quality and build a feature store used for machine learning? Please visualize your solution and explain the advantages and disadvantages.

- **Task 2: Infrastructure-as-Code (IaC):**
  - 1. Question: Why is IaC so important and what tools do you know to implement it on AWS? Can you explain the difference of those tools in a nutshell?

- 2. Question: Based on your solution in Task 1 (2nd Question), what might a Terraform implementation look like? You don't have to write code! It's enough if you show the structure of the code repo and explain what part of the infrastructure you define where.

- 3. Question: How can you define the infrastructure for a staging and production environment without duplicating the Terraform code?

You get the chance to present your solutions in our interview. Please send us your solution at least one day in advance.

Have fun with the tasks ☺