# CASE STUDY

DATA ENGINEER

# DATA ENGINEER
BACKGROUND

The purpose of the **Consumer Analytics (CA)** team is to deliver analytical products to adidas' Digital organization. These products may include reports, dashboards, data models, data pipelines as well as data science and machine learning algorithms.

Our **products** address mostly adidas-internal audience (e.g digital analysts, management, planning teams) but not only – some of them also build the foundation for consumer facing products like digital communications or web-site & app personalization. While our full delivery capacity is assigned to Digital, we organizationally sit within a different part of the organization, the Data & Analytics (DNA) department.

The CA team is internally staffed with ~15 people, distributed across the three functional areas: project management, architecture and data engineering. In addition to this, we are working a lot with external vendors, esp. on the data engineering side. To strengthen our internal footprint, we are now looking into growing the **internal data engineering team** to an overall total size of ~20. Growth is expected to happen mainly in our Tech Hub in Gurgaon/India.

In this context, we are currently looking for an experienced **Data Engineer** in Gurgaon, who will be working mostly hands-on on engineering our analytical products.

# DATA ENGINEER
## OUR CHALLENGE

At the moment, CA is utilizing on **two separated environments**, Exasol and the Big Data Platform (BDP).

**Exasol** is an in-memory database optimized for fast read access while the **BDP** is a cloud-based service that is implemented based on the Hadoop framework and Amazon EMR. It includes several pre-bundeled / adapted components and is operated by adidas Platform Engineering & Big Data teams.

Depending on the concrete environment, we currently use the following **tech stacks**:

- Exasol: mostly Alteryx for ETL supported by Lua (in Exasol), Alteryx Server for scheduling and orchestration, Python and R for data science models and cron for scheduling
- Big Data: mostly Spark/Scala for Data Integration & Transformations, querying of data through Hive and SparkSQL, supported by Zeppelin notebooks & R/Python (aka the "Big Data Lab")

We are in the midway of **re-shaping our architecture** to be ready for future challenges with the main ambition is to get rid of technology silos and to overcome its individual limitations.

In this context we already invested some efforts into **improving our architecture**, utilizing Bitbucket (git), Jenkins/Groovy and Docker/Kubernetes. As a next step we want to further standardize and move collectively more towards Python as programming language with different flavors/libraries with respect to the environment (e.g PySpark on the BDP, pandas).

We are searching for a **Data Engineer**, able to support us on this journey.

# DATA ENGINEER
## CASE STUDY - YOUR TASKS (1/2)

**Open Library** is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. In the section Bulk Data Dumps, they provide public feeds with the library data.
→ https://openlibrary.org/developers/dumps

They also provide a **shorter versions** of the file for developing or exploratory purposes, where the size is around 140MB of data instead of ~20GB of the original/full file (referring to the "complete dump").
→ https://s3-eu-west-1.amazonaws.com/csparkdata/ol_cdump.json

Starting with the short version of this file, pls. **download** it to your local laptop:

```
wget --continue https://s3-eu-west-1.amazonaws.com/csparkdata/ol_cdump.json -O
/tmp/ol_cdump.json
```

Note: This is an open exercise, you can use **whatever technology you might find useful** to showcase your solution, cloud services, on premise, etc. We suggest to provide also an architectural overview with a diagram that shows all the components together with a short explanation of each of them.

# DATA ENGINEER
## CASE STUDY - YOUR TASKS (2/2)

Please use the JSON file to provide the following information.

1. Load the **data**
2. Make sure your data set is **cleaned** enough, so we for example don't include in results with empty/null "titles" and/or "number of pages" is greater than 20 and "publishing year" is after 1950. State your filters clearly.
3. Run the following **queries** with the preprocessed/cleaned dataset:
    1. Select all "Harry Potter" books
    2. Get the book with the most pages
    3. Find the Top 5 authors with most written books (assuming author in first position in the array, "key" field and each row is a different book)
    4. Find the Top 5 genres with most books
    5. Get the avg. number of pages
    6. Per publish year, get the number of authors that published at least one book
4. How would you design a **scheduled data pipeline**, which would load this data on a daily basis? Please explain the design principles applied if any (and why).

You have a total of **45 minutes** to explain your approach to the case study to the interviewing panel, typically formed by the Senior Director of Consumer Analytics and his first line.

*Final note: this is not about being picture-perfect and presenting glossy marketing slides – ideas, solution quality and architecture count more than a nice-looking presentation!*