



CASE INTERVIEW

DATA ANALYST

SITUATION

The focus of the GAC Group in previous years was mainly on information governance with emphasis on master data management and data quality management. At the same time, GAC Group has been investing in data availability and data integration by building both a Data Lake and a Data Warehouse to prepare a healthy foundation for reporting, self-service BI, and data analytics.

From 2021 onwards the focus is shifting towards reporting, business intelligence, and data analytics. This is the reason why GAC Group IT is hiring a Data Analyst as part of Digital Product Management team. The first main task of a newly hired data analyst will be to help in identifying potential use cases in data analytics based on available data that could have the biggest impact on the business and start implementing those use cases.

Due to the confidentiality of GAC data, for the purpose of this interview you will be working with a prepared data set consisting of 1,022,141 sales transactions based on a fictive company spanning through two years 2010-2011. Company mainly sells unique all-occasion giftware. Each transaction consists of the following attributes:

- **Sales Transaction**
 - ID – unique transaction **identifier** [*integer*]
 - InvoiceNumber – unique document number associated with the sales invoice [*text*]
 - InvoiceDate – date and time of the transaction [*date & time*]
 - CustomerID – ID of the customer, references ID in customer data set [*integer*]
 - CountryID – ID of the country of the customer, references ID in the country data set [*integer*]
 - ProductID – ID of the product sold, references ID in the product data set [*integer*]
 - Price – sales price [*decimal*]
 - Quantity – sold quantity [*integer*]
 - Amount – sales amount calculated as sales price multiplied by quantity [*decimal*]

Data set is provided with three supporting master/reference data sets, including:

- **Customer** – 5843 customers in total
 - ID – unique identified of the customer [*integer*]
 - Code – unique customer code [*text*]
 - Name – customer name [*text*]
 - CountryID – ID of the customer country, references ID in country data set [*integer*]

- **Product** – 4582 products in total
 - ID – unique identifier of the product [*integer*]
 - Code – unique product code [*text*]
 - Name – product name [*text*]
- **Country** – 250 countries in total, customers coming from 40 countries
 - ID – unique identifier of the country [*integer*]
 - Code – unique country code [*text*]
 - Name – country name [*text*]

The goal of this interview is to:

1. Identify as many potential use cases in data analytics as possible based on provided data sets and specify which supervised or unsupervised learning approaches would be used for each use case. Data set provided allows use cases based on at least classification, regression, and clustering learning approaches but it is not limited to them.
2. Prioritise data analytics use cases starting with the one that could have the biggest impact on the business (primarily benefiting sales and marketing departments in this case) based on a prioritisation criteria defined by you.
3. Implement and demo a use case identified by you as top priority.

You are free to enrich the provided data sets in any way possible if you think it will add meaning to the case interview and your presentation even though enriching of the data is not a mandatory step.

TASK 1 – IDENTIFYING USE CASES BASED ON AVAILABLE DATA

Explain your data exploration process, data preparation process, and how did you go about identifying potential use cases based on available data? Which hypotheses can you form based on the data? List the identified data analytics use cases and identify appropriate supervised or unsupervised learning techniques that would be used to train your learning models for each use case. What potential inputs and outputs of each use case can you identify?

TASK 2 – PRIORITISING USE CASES

How would you prioritise the use cases identified in task 1? Which prioritisation criteria would you use? How would you estimate and measure the impact on the business of each use case? Which use case have you identified as top priority?

TASK 3 – IMPLEMENTING A TOP PRIORITY USE CASE

For identified top priority use case develop a learning model based on provided data. How did you select and test the input variables? How did you test the results? How reliable are the results? Visualise & demo the results by using any available data visualisation tool. How would you productionalise this use case?

INTERVIEW REQUIREMENTS

- Presentation of your choice addressing the three tasks above
- Demo of the solution for the prioritised use case
- One-page executive summary as a supporting document
- More detailed report as a supporting document

IMPORTANT CONSIDERATIONS

- Consider the audience to be CEO, CIO, Directors and Managers
- You will have 30 min to present and 15-30 min for Q&A and open discussion
- You can enrich the provided data set if required
- You can ask for clarifications before the interview by emailing questions to josip.nadih@gac.com

EVALUATION CRITERIA

- Logical reasoning
- Creativity of the approach
- Problem-solving skills
- Technical skills
- Documenting skills and language
- Presentation skills and language