You are a Data Scientist working for a global leader in the FMCG market. Your task is to analyze sales data from multiple sources and formats and create a pipeline which will store the data in one place. Those can be used to predict sales future volumes. Your work will help senior management in preparing warehousing strategy as well as optimize supply chain management. There are high savings possible to achieve and senior management is looking closely at your work. You will be granted a 10 minutes time slot where you will have to present the solution. Please note that your are expected to cover the following areas:

* Data your were able to gather
* Technical description of the solution, not so much detailed, high level only
* Anticipated costs of the solution
* Possible improvements in future/new features

# Data Sources

The data from different sources can be uploaded to Azure Data Lake Storage Gen2 easily, so this is your starting point. Files are in different formats. You were able to identify the following:

* 1. CSV
  2. Apache Parquet
  3. JSON

Your pipeline needs to be able to ingest those data and aggregate them in one place.

# Pipelines

It is up to you which solution will be used. (Azure Synapse, Azure Databricks or ADF

- all of those were discussed during the training). Please note you must present reasons for choosing the particular solution and costs behind it.

# Database

You can use any database that will hold all data.

1. (Bonus Points) **Streamlit** is one of the most recent python libraries for powerful dashboard creation and an open source framework for machine learning and data science teams. It can produce industry standard and visually attractive dashboards with very simple coding.

Visual Presentation and live dashboard which is easily accessible will definitely be useful for the management.

# Final notes:

All Data Sources are available under the specified location - link will be given to project members (project coordinator needs to create ADSL Gen2 with the files and share access.

You can work in groups

1. Identify data and prepare high level assumptions on the workload, frequency etc. (4pt)
2. Choose high level solution and prepare business drivers (6pt)
3. Create CSV pipeline (2pt)
4. Create JSON pipeline(2pt)
5. Create Parquet pipeline (2pt)
6. (Bonus points) Connect Streamlit to data source (3pt)
7. (Bonus points) Present Dashboard with the data (3pt)

# Possible Solution High Level Architecture:

Please note that this is one of the many solutions:

