**Oliver Wyman Online Case Study: Data Engineer Profile**

**Background:**

DMU Current pipelines utilizes AWS cloud infrastructure EMR+S3+EC2 and spark ML pipelines for Modelling and Scoring on Big Data. On daily basis we process 3 billion rows of transactional data.

**Problem Statement 1:**

Data: Attached Dataset “Final\_model\_data” has sample model data in tab “modeldata”, description of this data is given in tab “data\_description”

Exercise:

1. Expand this data 100 folds by either duplicating the rows or by using some logic for Data Expansion.
2. Create a Spark Schema on this data set.
3. Predict Target variable using all other independent variables:
   1. Idea is not to get accurate model but test your coding skills.
   2. Design Pyspark ML pipeline for both scoring and Modelling.
4. Provide a clearly commented code and a readme file sharing your logic and steps undertaken.
5. Zip entire solution (Code and Readme file) in problem\_statement\_1.zip and share via email at [panchsheel.mittal@oliverwyman.com](mailto:panchsheel.mittal@oliverwyman.com)