Task: Adding fields to financial elimination dimension

The following fields needs to be added to the dimension: SourceType, Company

- **Field SourceType**

The source type field is one that is only present for BC rows. That also means that duplicates would be generated for some specific rows. One of the ways we dealt with that was to create the following transformations. The source type name is the column that only exists in BC and therefore for all AS4 rows it would be litted as *null.* For example there could be an internal transaction from one company to the other both from AS4 and BC with the same elimination number which constitutes the primary key. The problematic thing about that is, that we need to have unique primary keys. In order to make sure we handle this, what we do is to

1. select from the *set* of columns removing the "SourceTypeName".
2. Group the data frame by the set of columns
3. Use collect_list on SourceType name to get the source type name as a list data structure
4. Create column that selects the lits "Multiple Types" which should not be possible, otherwise takes the first element.

The smart thing about this solution is that when we collect list, it removes null values thus if there are both an AS4 row and BC row it will only take the BC row.

```
cols = set(df_as4.columns) - {'SourceTypeName'}

 df_res = (df_as4.union(df_bc)

                  .withColumn('SourceTypeName', when(col('SourceTypeName') == ' ',
lit(None)).otherwise(col('SourceTypeName')))

                  .groupBy(list(cols))

                  .agg(collect_list('SourceTypeName').alias('SourceTypeName'))

                  .select(df_as4.columns)

                  .withColumn('SourceTypeName',

                        when(size(col('SourceTypeName')) > 1, lit('Multiple Types'))

            .otherwise(col('SourceTypeName')[0]).alias('SourceTypeName'))

      )
```

- **Company**

The company is a field that shows which company is involved in the internal transaction. To extract that one can do some substring transformations such as the following:

```
withColumn('Company',
                    when(substring(col('EliminationName'), 0, 12) == 'Intern andet',
                       substring(col("EliminationName"), 14, length(col("EliminationName"))))
                    .when(substring(col('EliminationName'), 0, 10) == 'Intern køb',
                        substring(col('EliminationName'), 12, length(col("EliminationName"))))
                    .when(substring(col('EliminationName'), 0, 11) == 'Intern salg',
                        substring(col('EliminationName'), 13, length(col("EliminationName"))))
                    .otherwise(col("EliminationName"))))
```