Analytics mindset

ETL

Case 2 – Text extraction and unique identifiers – Alteryx

In older computer systems, multiple values were often stored in a single cell to save space. This practice is sometimes still followed today. For example, an employee identification number may tell you the employee number, plant number and business function. That is, 143-01-Acc could identify employee 143 from plant 01, who works in accounting.

For this case, you are provided with an Excel data file, named **Analytics_mindset_case_studies_Case2_Alteryx.xlsx,** that contains 597 rows of employee data. In the tab labeled Case 2 data, you will find three columns: EmployeeCode, FirstName and LastName. The EmployeeCode is the combination of four different fields: Location, EmpID, PlantID and PayPeriod. Each of these fields is defined as follows:

► Location: The location code shows the location where the employee works. The company operates in eight different countries: Argentina (ARG), Australia (AUS), Canada (CAN), England (ENG), Germany (GER), Japan (JAP), Mexico (MEX) and the United States of America (USA). The country codes are always three digits and are the first three digits in the EmployeeCode, reading from left to right.

► EmpID: The company assigns a random employee identification number from 1,000 to 1,597. Reading the EmployeeCode from left to right, the EmpID is the first set of numbers immediately after the Location code and preceding the dash.

► PlantID: The company has various plants throughout the different countries. Each country numbers its plants starting at 10, and adds one more number for each additional plant. The PlantID is contained in the EmployeeCode, reading from left to right, immediately after the dash.

► PayPeriod: Employees are paid either weekly or monthly. The system records this as a W for weekly and as an M for monthly. The PayPeriod is the last letter, reading from left to right in the EmployeeCode.

You have been asked by your manager to extract data using the employee code and also to create a new unique identifier that will provide the plant number by location.

Required

► Complete the ETL overview case, which covers the fundamental considerations in the ETL process.

► Use Alteryx to:

  – Create separate columns for each of the four fields described above. The field labels and the number and type of characters for each field are listed below:

    ► Location (three alpha characters)

- ► EmpID (four numeric characters)

- ► PlantID (two numeric characters)

- ► PayPeriod (one alpha character)

– Add a concatenated field to create the unique identifier combining Location and PlantID so the output looks like the following: USA-12

► Produce an output file in Excel. Submit your Alteryx workflow as a packaged workflow (.yxzp file type [Options > Export Workflow >]) and include your name in the file name.