Laurel Technology Solutions Ltd.

Data Standardization Plans and Challanges

Goals

Laurel Technology Solutions Ltd., has recently begun to utilise the data they obtain from their customers. They do not currently have a single cohesive record representing all of their customer data. They are looking to unify these ahead of further data investigation and exploitation, and to pool these data together into a central database.

In line with the objective, the goals are to

- Extract data from various given datatypes
- Transform the data into a consistent datatype and format
- Load the data into a database

Given Data and their Formats

Three formats of data that contained useful information were given. JSON, XML and CSV. Each of these data formats have the ability to store information in an almost tablelike organized manner

XML

Attributes Data

Sample

```
<user firstName="Iain" lastName="Dixon" age="64" sex="Male" retired="False" dependants="2" marital_status="single" salary="56185" pension="0" company="Hudson PLC" commute_distance="14.1" address_postcode="G2J 0FH"/>
```

JSON

Banking Data

Sample

```
{"firstName": "Nicholas", "lastName": "Spencer", "age": 71, "iban": "GB43YKET96816855547287", "credit_card_number": "2221597849919620", "credit_card_security_code": "646", "credit_card_start_date": "03/18", "credit_card_end_date": "06/26", "address_main": "462 Marilyn radial", "address_city": "Lynneton", "address_postcode": "W4 0GW"}
```

CSV

Vehicle Data

Sample

```
First Name,Second Name,Age (Years),Sex,Vehicle Make,Vehicle Model,Vehicle Year,Vehicle Type
Lynne,Hudson,37,Female,Subaru,Forte,1993,Convertible
```

The challenges experienced in combining the various data formats

✓ Different key values for the same information

The column header for columns that have the same information in other formats in the CSV format was quite different. The column header made use of capitalizing each word while the other format used a mixture of camel and snake case. To solve this issue, I had to rename the column header name of the csv file.

✓ Converting all to one data format was difficult

The data is given in XML, JSON and CSV format and provides some difficulty in converting them to a unified format. Python interprets the formats differently. To parse XML format, a python library is used to get the attributes from the tree-like structured data into a python list. The JSON format was interpreted as a python dictionary with key and value pairs. Because of the differences between the methods of dictionaries and lists, it was difficult to process the data in a similar format. The solution I had to this was to leave the JSON format as is so that it could be easily inserted into a temporary database for processing.

✓ Other challenges would be seen on the python script file

Difficulties that may be encountered by Laurel Technology

Challanges

- Collecting Data in different format could lead to complications and cause confusion
- Lack of customer unique identifier could make joining the table to other tables difficult
- The current database solution may not be adequate for rapid expansion in the long run

Possible Solution

- If possible, all data should be collected in one format, preferrably CSV or JSON format. That way, the ETL process would be seamless
- Each customer should be given a unique identifier that would differentiate customers and prevent errors while transforming the data. It would also help to join with other tables such as the sales or order table
- In order to plan for more expansion in the future, Laurel Technology should consider putting a more robust system of data collection and transformation that would perform task automatically with little human intervention

Thank you!!