Project Thales

Introduction

In lead generation industry, many produce lots of semi-structured data, mostly in CSV and Excel Spreadsheet, but they are not able to organize them into proper structured data. As part of “Project Thales”, we should build tools and app that helps to convert building structured database and “simple” frontend application to manage the data.

Background

Most of the lead data are in excel, and historically we have 1000s of spreadsheet containing lead data. In order to convert them into proper DB, we should first understand the problem.

There are two kinds of details are present, actual data inside the file, and file metadata itself would give additional clue about the data. For an example

Name	**Designation**	**Salutation**	**Company**	**Email**	**Address**
Aravind	CEO	Mr.	IBM	Aravind@ibm.com	Chennai
Bhavna	MD	Mrs.	Google	Bhav@google.com	Bangalore

If we have information something like above from a file that was last modified in 2012, Oct-15, then all the data were true as on day 15-Oct-2012, Very next day lead data may not be valid, hence while processing the data, we should be careful enough to retain some of the metadata information

Problem statement

1. How to process 1000s of file that has above information and create a valid and normalized database?
2. How to find duplicate records that spil across multiple file?
3. Once all the processing done, can we refresh using one more new file find on new disk?

<<Fill in your solution>>

- List unique files that are having .xls, .xlsx
    - Here two file can have different name but it may have same content

- Ignore the filename, but create unique key using following combination
    - File_extension, md5 checksum of the file
    - So a file “IBM_lead_gen.xls” would have a key “xls_394939439493.xls”
- Create a meta_data table with following column name
    - Source
    - Sheet
    - TableName
    - DateCreated
    - DateUpdated
    - ColumnNames
        - Only one column for all the column name in excel, concatenate using comma
- Once unique file keys are identified, we have to dump them into a database
- For each file, choose only first sheet of the spread_sheet
    - Read the tabular data, and create table with the set-of-column found in the first column of the row
    - Insert all the records
    - Here table name could be be either Filename_md5
    - Update meta_data table with all the details of the file
- Now we should have 1000s of table + 1 table that has all the meta-data

MD5

Creates unique checksum based on content of the file

Technical specification