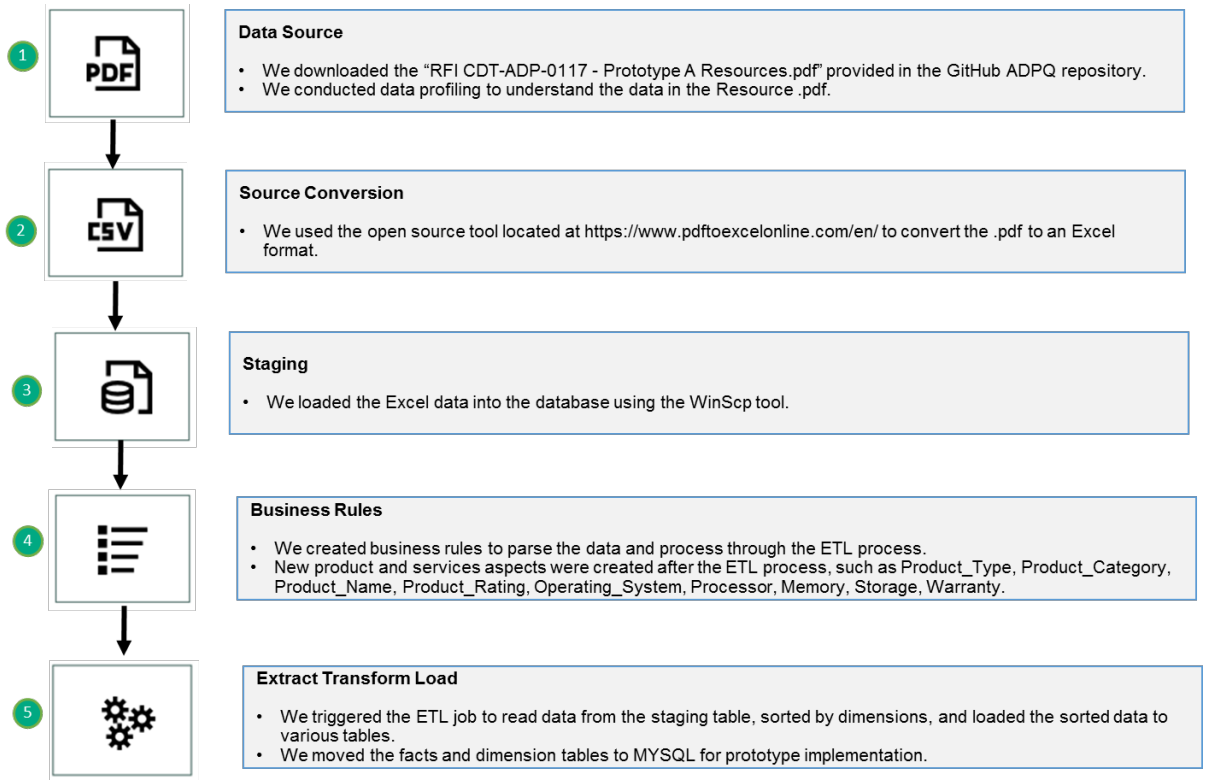
ETL Process for Cal eStore Prototype

The following graphic depicts the Extract Transform Load (ETL) process our team utilized in developing the Cal eStore prototype.

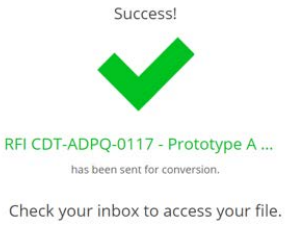


The following narrative and graphical depictions show the steps in more detail.

Hewlett Packard Enterprise

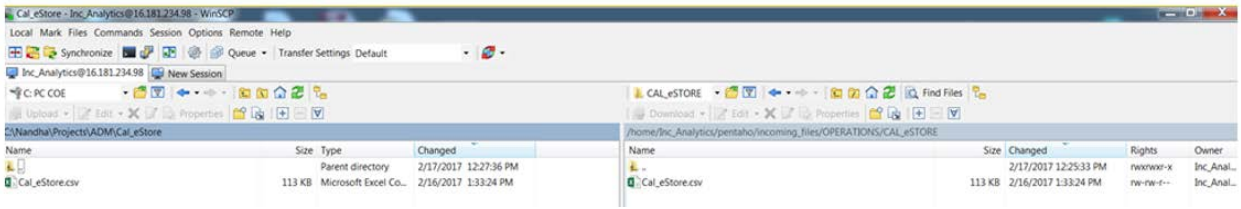
Data Conversion

RFI CDT-ADPQ-0117 - Prototype A Resources.pdf provided in GitHub for prototype A was not a MYSQL DB (database) consumable format; therefore we used an open source PDF to Excel conversion tool. This tool can be found at https://www.pdftoexcelonline.com/en/ — to extract the data and convert it to a .csv file.



Data Loading

We converted the input Excel file into .csv format, then moved the file to a server drop zone where it was loaded into the database using the WinScp tool.



Business Rules

The Prototype A use case requires a comparison of products and services, as well as grouping into high-level categories like Hardware, Software and Services. Data in the *RFI CDT-ADPQ-0117 - Prototype A Resources.pdf* file is ***not*** conducive for item comparison and application layout. For example, hardware product aspects like product name, operating system, processor, memory, storage, warranty, connectivity type, HDD RPM, and RAM type, were merged together in a single column “Item description.” The same issue was there for all other products and services in the resource .pdf file.

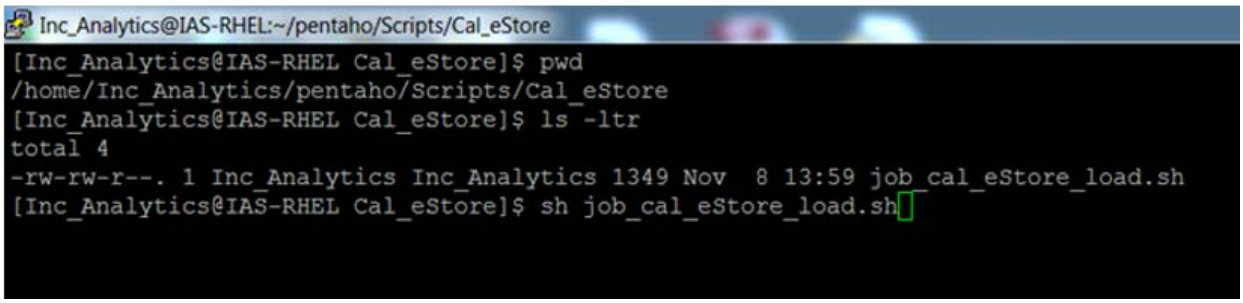
To efficiently parse the item description text, we created an Extract, Transform and Load (ETL) process. The ETL process required development of business rules to ensure consistent parsing of the data.

NOTE: The following example is for illustration purposes and is not a complete list of business rules.

Business Rule Objective	Description	Applicable Column	Rank
Divide each records in to three Product Types (Hardware, Software and Services) and Product Category	BR1. When the Horizontal Category equal to Configuration (Hardware) and Category equal to (Standard Desktop Hardware or Workstation Hardware or Thin Client Hardware or Power Desktop Hardware or All in One Hardware) then Product type equal to “Hardware” and Product Category equal to "Desktop"	Product_Type Product_Category	1
Divide each records in to three Product Types (Hardware, Software and Services) and Product Category	When the item description contains "DDPE" then Product type equal to “Software” and Product Category equal to "Date Encryption" When the item description contains "Computrace" and then Product type equal to “Software” and Product Category equal to "Device & Data Security" When the item description contains "Intel Standard Manageability" or "INTEL VPRO"" and then Product type equal to “Software” and Product Category equal to "Others"	Product_Type Product_Category	2
Divide each records in to three Product Types (Hardware, Software and Services)	BR2. When the Item Description contains = (FHD and LCD or LED and HD or Widescreen or Touchscreen or EliteDisplay or LED and FHD or 4K and LCD) and Sub_Category_Detail (Upgrades or Upgrade or MONITOR) then Product type equal to “Hardware” and Product Category equal to "Monitor"	Product_Type Product_Category	3
Create the product name for each record	BR3. Product name equal to "Manufacturer" + first part of the "Item Description"	Product_Name	4
Divide each records in to three Product Types (Hardware, Software and Services) and Product Category	BR4. When the Horizontal Category equal to Configuration (Hardware) and Category equal to (Integrated Microphone or External Speakers or Business Speakers or Wireless Headset or Stereo Speakers or LCD Speakers) then Product type equal to “Hardware” and Product Category equal to "Audio Devices"	Product_Type Product_Category	5
Divide each records in to three Product Types (Hardware, Software and Services) and Product Category	BR5. When the Horizontal Category equal to Configuration (Hardware) and Category equal to (DVD ROM or External Drive or Optical Disk Drive or USB DVDRW Drive) then Product type equal to “Hardware” and Product Category equal to "DVD Drive"	Product_Type Product_Category	6
Divide each records in to three Product Types (Hardware, Software and Services) and Product Category	BR6. When the Horizontal Category equal to Configuration (Hardware) and Category equal to (NVIDIA Quadro or AMD Radeon or AMD Graphics card or AMD R5) then Product type equal to “Hardware” and Product Category equal to "Graphics Card"	Product_Type Product_Category	7
Divide each records in to three Product Types (Hardware, Software and Services) and Product Category	BR7. When the Horizontal Category equal to Configuration (Hardware) and Category equal to (Solid State Drive or Hard disk Drive or Optical Disk Drive) then Product type equal to “Hardware” and Product Category equal to "Hard Drive"	Product_Type Product_Category	8
Create the Display Attributes for comparison	BR8. When the Product type equal to “Hardware” and Product Category equal to "Monitor" then move the numerical (14" or 15" or 15.6" or 17.3" or 19" or 19.5" or 21" or 22" or 23" or 24" or 28" or 30") in the Item Description to a new attribute Disply_Size	Display Size	9
Create the Display Attributes for comparison	BR9. When the Product type equal to “Hardware” and Product Category equal to "Monitor" then move the (LCD or LED) in the Item Description to a new attribute Disply_Type	Display_Type	10
Create the Display Attributes for comparison	BR10. When the Product type equal to “Hardware” and Product Category equal to "Monitor" then move the (4K or FHD or HD) in the Item Description to a new attribute Display_Definition_Type	Display_Definition_Type	11
Create the Warranty Attributes for comparison	BR10. When the in the Item Description contain the word Warranty then move the [(Numeric and year description) and (Lifetime)] to a new attribute "Warranty"	Warranty	12

ETL Job Steps

We triggered the shell script ‘job_cal_eStore_load.sh’ to load the data through the ETL job.

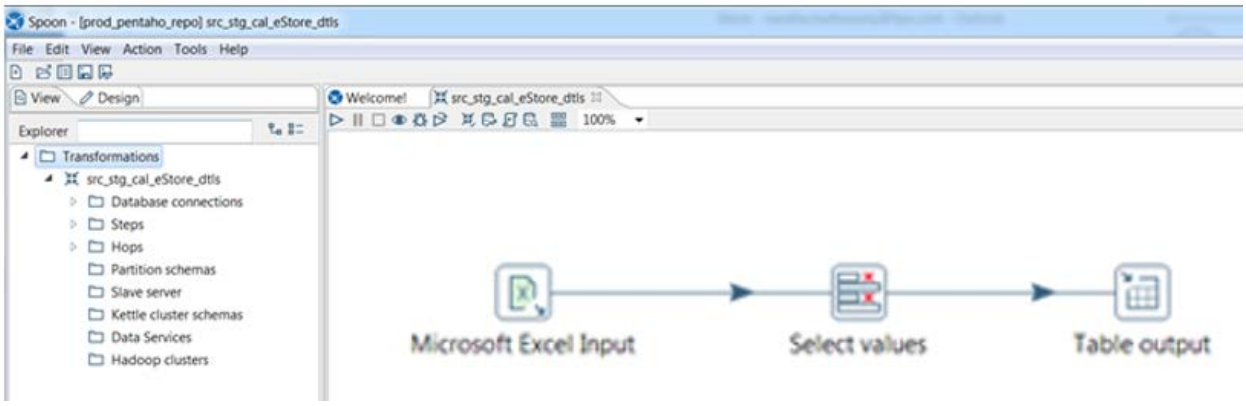


The following depictions describe the individual steps in the ETL process.

Hewlett Packard Enterprise

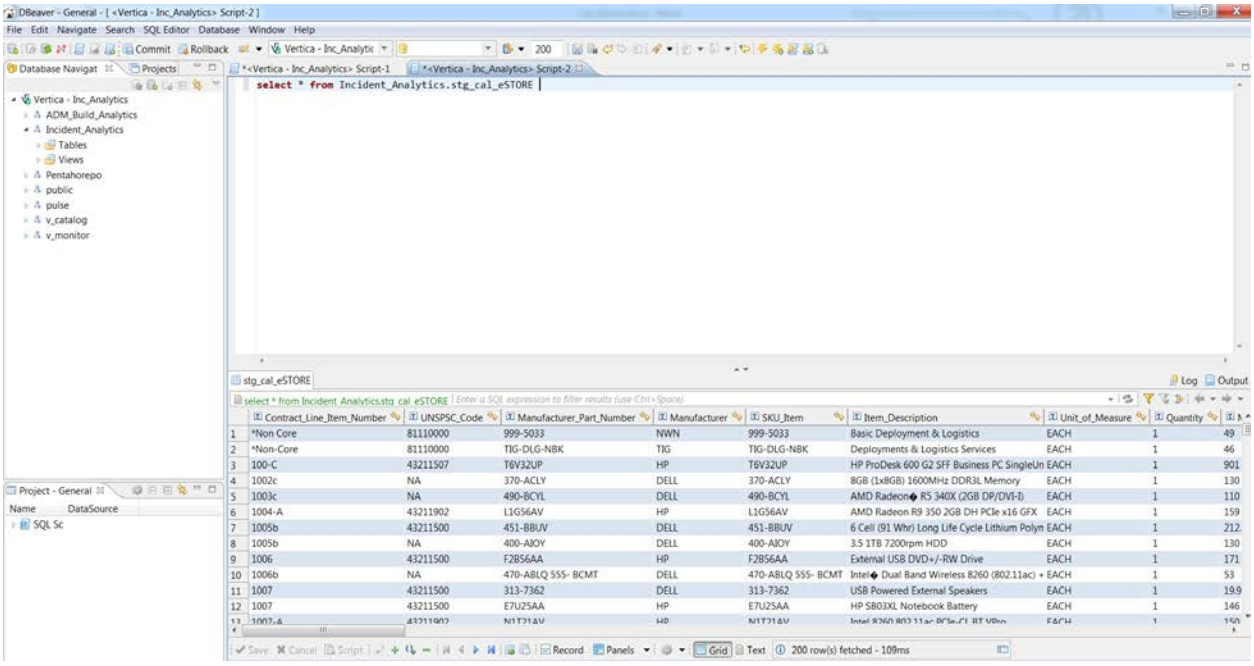
Step 1 — ETL staging job

This job reads data from .csv file and loads it into a staging table.



Step 2 — Staging table

Data was successfully loaded into the staging table ‘stg_Cal_eStore’.

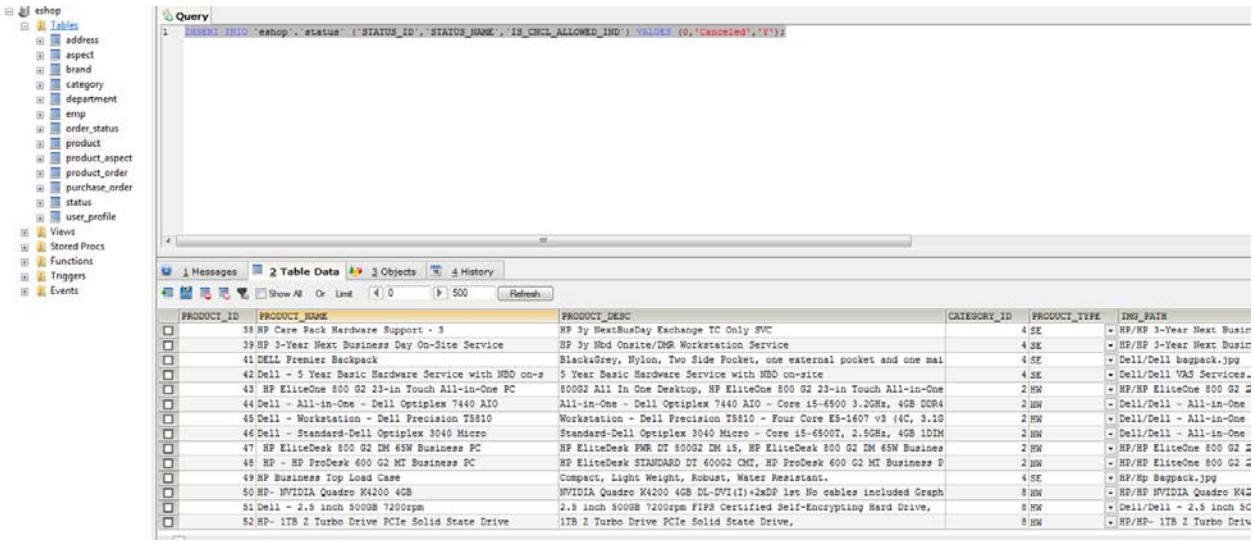


Step 3 — Facts and Dimension tables for Cal eStore

The ETL job in the following example reads data from the staging table, sorts by dimensions, and loads the sorted data to various dimension tables.

Hewlett Packard Enterprise

Records for Cal eStore were successfully loaded into dimension tables as illustrated in the following example.



Hewlett Packard Enterprise