Data Pipeline Document
29 November 2017

Team: Kids Helping Kids on 45th

Team Members:
Gary Gregg
Abhishek Varma
Jahnavi Jasti

What are the data streams that you are using?

Kids on 45th needs work with the legacy sales data in order to better understand item pricing. Their legacy sales data is stored in four MS Access databases (*.mdb) extension. The databases are named *Custdata.mdb*, *Product.mdb*, *Sales.mdb* and *Scandata.mdb*.

The **Custdata** database has seven tables: ‘Users’, ‘Sales’, ‘Maillist Profiles’, ‘Hold’, ‘Customers_Finance’, ‘Customers’, and ‘Cnwa’.

The **Product** database has six tables: ‘ProductsBu04Older’, ‘PP_Presets’, ‘PP_Descriptions’, ‘PP_Catagories’, ‘PP_Active_Products’, and ‘Active Back-Up 3600’.

The **Sales** database has seven tables: ‘Transaction Types’, ‘State Sales Tax’, ‘Sales’, ‘PP_Soldprod2008’, ‘PP_Sold_Products’, ‘PP_Returned_Products’, and ‘Gift_Cirt’.

The **Scandata** database has two tables: ‘CC_Descriptions’, and ‘CC_Con Check In’.

Note: The structure, meaning, and organization of the data is still under analysis. This documents only identifies sources and data point counts.

How many data points do you have?

Record counts for tables in the **Custdata** database:

Table Name	Record Count
Users	10
Sales	209,878
Maillist Profiles	1
Hold	1
Customers_Finance	8,793
Customers	8,816
Cnwa	1,904

Record counts for tables in the **Product** database:

Table Name	Record Count
ProductsBu04Older	29,223
PP_Presets	279
PP_Descriptions	4
PP_Categories	5
PP_Active_Products	147,563
Active Back-Up 3600	17,941

Record counts for tables in the **Sales** database:

Table Name	Record Count
Transaction Types	6
State Sales Tax	3
Sales	210,819
PP_Soldprod2008	76,587
PP_Sold_Products	538,024
PP_Returned_Products	0
Gift_Cirt	0

Record counts for tables in the **Scandata** database:

Table Name	Record Count
CC_Descriptions	19
CC_Con Check In	106,303

Where does the data reside?

The legacy sales data has been shared with Kids Helping Kids on 45th team members by store owner Bookis Worthy using **Dropbox** and associated technology. The legacy sales data is not currently being updated, and we can safely assume the data is in its final form. The data has been downloaded to a team member laptop in order to perform the analysis required for this document.

What software are you using to access the data?

Abhishek and Jahnavi: Gary has installed GNOME MDB Viewer on my Ubuntu Linux laptop in order to view the legacy sales data, and count the existing data points. If this application proves inadequate for the purposes of analysis, he will use appropriate technology for MacOS to view and access these files. If you are using Windows, please note here that you will be using MS Access to work with the data. If you are using MacOS, please research and determine what Apple technologies exist for working with MS Access datafiles.