Getting down to business:
Whiskey Business

Dylan, Chris and Patrick

Refresh: The Problem

Copyright © Free Vector Maps.com

■ Control States
★ Control Counties
- Montgomery County, MD
- Worcester County, MD
- Wilcomico County, MD
- Somerset County, MD
★ Some Jurisdictions in Alaska
Minnesota & South Dakota

Data: Getting and storing

1. Download data from 4 sources using Python:
   a. Virginia liquor price list (csv)
   b. Reddit Whisky Archive (csv)
   c. Metacritic whisky analysis (html scrape)
   d. Proof 66 price list (html scrape)

2. Clean and Transform data using Python:
   a. Filter out non whisky data from VA price list
   b. Generate a key based on brand name across 4 data sources
   c. Convert currency to US dollars
   d. Convert metric units to ounces
   e. Output as 4 csv files

3. Load and transform data into HDFS and Hive:
   a. Put raw csv files into HDFS
   b. Generate DDL based on columns and file names
   c. Run SQL to copy data from HDFS to Hive

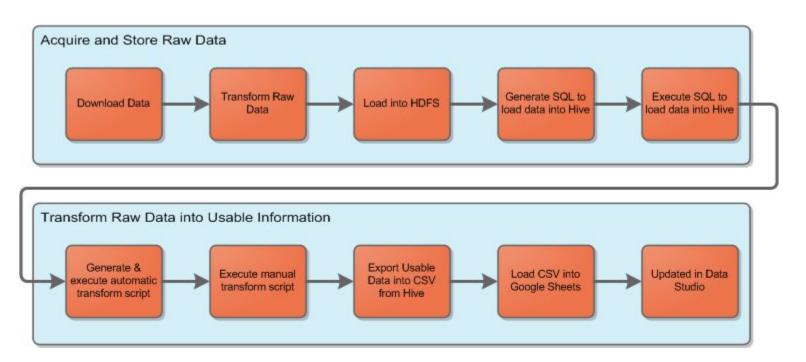# Data: Getting and storing (continued)

4. Transform hive data and join into one master table
   a. Transform numeric data into floats
   b. Join tables on key brand name
   c. Avg scores, prices, calculate price/oz

5. Export data to CSV and Google Sheets
   a. Use Hive to export master table to csv file
   b. Use python to upload csv file into Google Sheets

6. Connect Data Studio to Google Sheets

Google Sheets

Google Data Studio

The Big Picture

Acquire and Store Raw Data

Download Data → Transform Raw Data → Load into HDFS → Generate SQL to load data into Hive → Execute SQL to load data into Hive

Transform Raw Data into Usable Information

Generate & execute automatic transform script → Execute manual transform script → Export Usable Data into CSV from Hive → Load CSV into Google Sheets → Updated in Data Studio

WHISKY BUSINESS

Check it out!

https://goo.gl/WhtU4q

Note: Must be logged in as your @berkeley account

What's Next?

- Current limitations
  - Data is limited to VA
  - Match rate across data sets is low
  - Reviews easily influenced by one person
  - Data is updated on demand and subject to data sources changing

- Evolving the project:
  - More data sources to increase usefulness to other states
  - Add more out of state pricing data (with locations)
  - Improve text processing for match rates
  - Show number of reviews with score and popular items to review
  - Transform review scores to be on the same scale
  - Connect Data studio to Hive (one day?)
  - Convert data studio project into a website

Closing Time

270+ Unique bottles paired with reviews

195+ Bottles with price benchmarks

One central location for Virginia buyers to research whiskey before purchasing