Data Engineering using Transit Data

Bruce Irvin

C-TRAN

Portland State University

Data Science

80%: Data Wrangling

Source: Kandel, Paepcke, Hellerstein, and Heer. 2011.] SIGCHI Conference on Human Factors in Computing Systems (CHI '11).

Data Science

80%: Data Wrangling

20%: Complaining about Data Wrangling

C-TRAN

Portland State UNIVERSITY

Data Engineering

Data-related activities separate from analysis and presentation

Data Engineering

Search, Find, Evaluate, Synthesize, Acquire, Get Permission, Communicate, Transport, Import, Ship, Store, Share, Compress, Encrypt, Secure, Explore, Familiarize, Transform, Resample, Clean, Validate, Fill in missing data, Remove Outliers, Smooth, Simplify, Remove Bias, Shape, Combine, Integrate, Cross Reference, Process, Distribute, Version, Curate, Update, Manage Access, Document, Annotate, Audit, Archive, Destroy

Source: these terms were taken from several papers referenced by the Kandel, et. al. SIGCHI ‘11 paper

Data Engineering

June 16 Portland-area Job Posting (monster.com):

> "...[we] offer a wonderful opportunity to perform, excel and grow in the insurance industry as a Data Engineer within the IT team! We are seeking a highly motivated individual with a successful track record in building automated data pipelines ..."

CS410/510: Data Engineering

New course at Portland State University (Winter 2021)

First class: 15 undergraduate students, 32 graduate students

People who made it happen: Natalie Leon Guerrero, Bryttanie House, David Post, Aman Singh Solanki, David Crout (C-Tran), Taylor Eidt (C-Tran)

Term Project

Build real data pipelines for Transit Data

Data Source: one month of C-Tran breadcrumb data

Data Source: same month of C-Tran CAD/AVL stop event data

Term Project

Part 1: Gather and transport C-Tran breadcrumb data

Part 2: Validate, transform and load breadcrumb data to DBMS

Part 3: 2nd pipeline for C-Tran CAD/AVL stop data and integrate

Part 4: Produce a video showing the results

Term Project: PSU’s Contributions

PSU provided:

1. Two web sites that hosted the breadcrumbs as JSON and the stop data as nasty HTML tables. (Aman Singh Solanki)
2. In class programming exercises to help students learn
3. Lots of troubleshooting support (David Post)
4. A visualization module to display bus speeds on a map

C-TRAN

Portland State UNIVERSITY

Term Project: C-Tran's Contributions

C-Tran and PORTAL provided:

1. One month of breadcrumb data (> 100 million data points)
2. Same month of CAD/AVL stop data
3. Presentation about C-Tran data systems (David Crout and Taylor)
4. Ongoing Q&A support
5. Celebrity Guest Judges! (David Crout and Taylor)

C-TRAN

Portland State UNIVERSITY

Term Project: Google's Contributions

Google provided a $50 Google Cloud Platform credit to each student.

(plus more if/when they used up their initial credit)

Google also provided a great, reliable, performant "Compute Engine" platform.

Term Project: Part 1 (of 4)

A. Create, configure Google Cloud Platform (GCP) linux virtual machine.
B. Develop a simple python program to gather breadcrumb data from website
C. Configure your VM to run your gathering client to run daily.
D. Allocate and configure a Kafka message passing “topic” at confluence.com
E. Enhance client to parse the breadcrumb data to produce individual JSON records.
F. Enhance client to send the individual breadcrumb records to your Kafka topic
G. Develop a python program to consume the breadcrumb readings from the Kafka topic.
H. Configure your VM to run your Kafka consumer daily for rest of the term.

Term Project: Part 2 (of 4)

A. research the contents/meaning of the breadcrumb data

B. design a database schema for the breadcrumb data

C. enhance your pipeline

1. validate the data
2. transform the data
3. install and configure a PostgreSQL database server
3. load the data into the DB server

F. validate the running pipeline with simple summary queries to the DB

C-TRAN

Portland State UNIVERSITY

Term Project: Part 3 (of 4)

A. access the CAD/AVL stop event data

B. enhance the DB schema for the stop event data

C. build a new pipeline for the stop event data

D. integrate the stop event data with the breadcrumb data

E. testing

Term Project: Part 4 (of 4)

Create a 10 minute video

A. Describe your pipelines
B. Use the provided visualization module to analyze your data

In addition to grading the videos we ran a fun competition and awards ceremony with Natalie, Bryttanie, DaveC and Taylor as celebrity guest judges.

Example Results (links to student videos)

Pipeline Architecture (James Fotheringham): link to video

Summary of Data (Andrew Wiles): link to video

Data Analysis (Ebele Esimai and Sara Alotaibi): link to video

Animated Visualization (Daniel DuPriest): link to video

Overall Best in Class (Genevieve LaLonde): link to video

Takeaways

real **Transit Data** brings Data Engineering education to life

The students learned many valuable skills

Modern software infrastructure enables complex projects

Students succeeding in their job searches

Next up for Data Engineering Course

More data

Additional analysis

Improve course structure

Include more topics