Parsing User Input for Database Normalization

DYNA ’26 – 04/27/2026

Mark Wolfson and Kori Smith

BIG

Who is BIG?

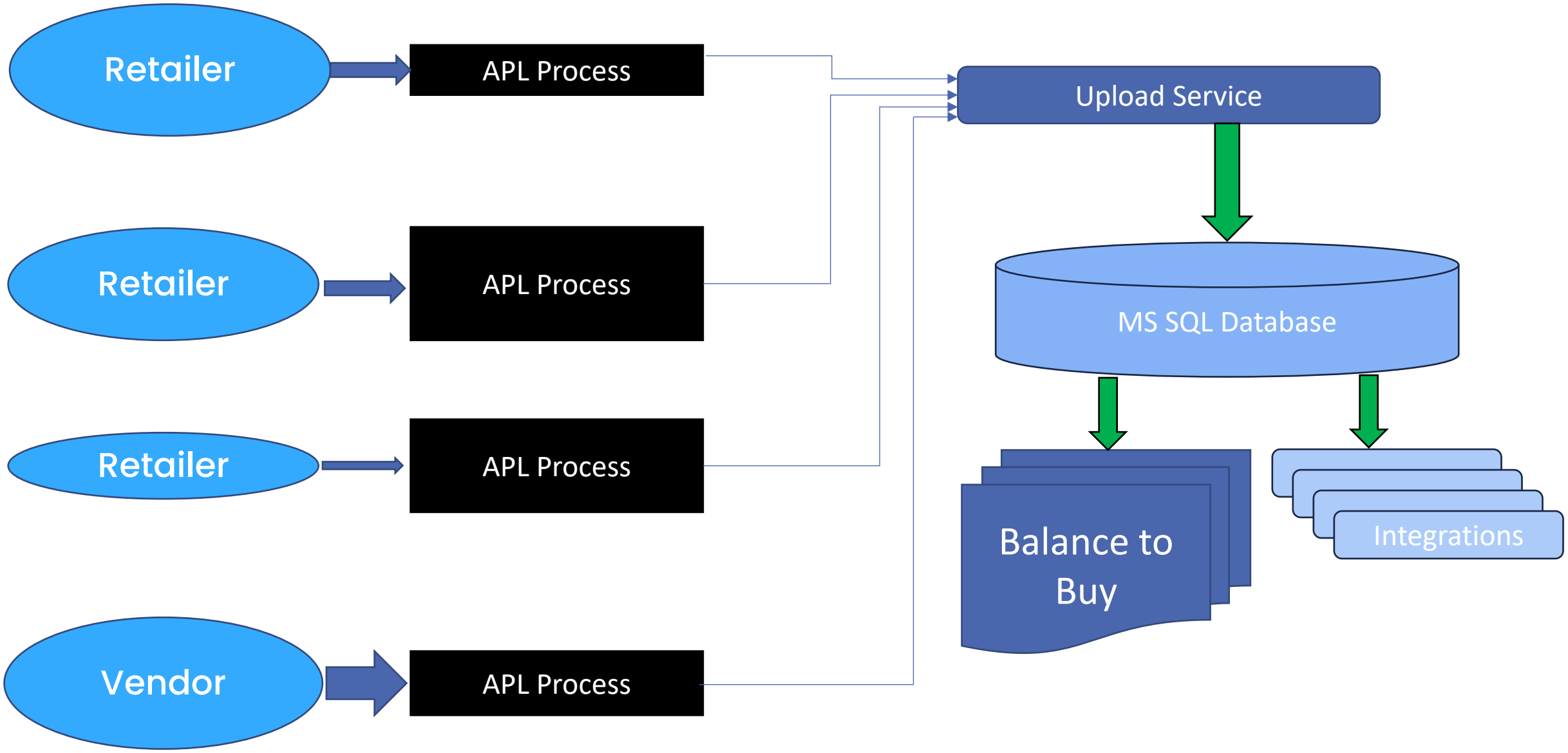
Who is BIG and why is it a perfect use-case for APL

- Service consultancy in the jewelry industry
  - Customers:
    - Retail Jewelers
    - Jewelry Manufacturers (vendors)
    - Jewelry industry service companies
  - Services:
    - Inventory analysis & merchandising consulting
    - Jewelry sales/inventory data aggregation
  - We provide these services by collecting accurate, timely data from 1600 retail jewelry stores using 40 different software systems, and 120 vendors.

BIG

Retailer Data – A HOT MESS

- Integration with dozens of systems
  - Varying number of files even for the same system
  - Different formats
    - Flavors of CSV
    - Text files
    - Excel
    - XML
    - JSON
  - Different content
    - Variations in content between retailers using the same system
    - Variations within a retailer depending on who's entering the data
    - Changes over time


Retailer
APL Process
Retailer
APL Process
Retailer
APL Process
Vendor
APL Process
Upload Service
MS SQL Database
Balance to Buy
Integrations

Retailer

.CSV

Applet
15 POS Systems

Retailer .csv API

Retailer .txt API

Retailer .xls API

Retailer .xlsx API

Vendor .xml API

Vendor .json API

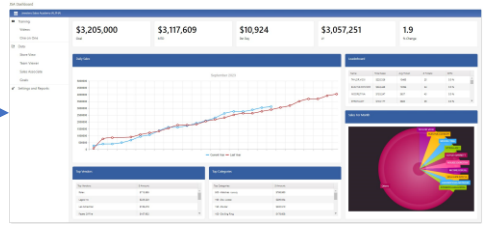
Raw Data Store

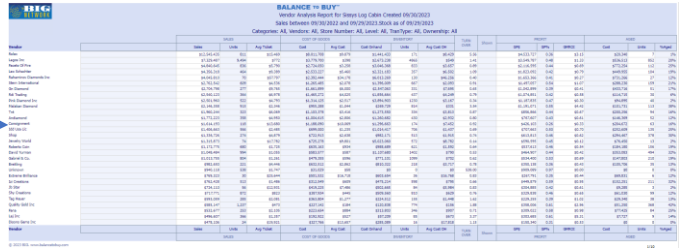
Retailer
Retailer
Vendor
Vendor
Retailer
Retailer

FTP

Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process
Black Box Process

MS SQL DB





Custom Reporting

Outbound Integrations

Outbound Datafeeds

Example descriptions

3.08Ct E VS1 Princess Lab Grown Diamond Cert Name Cert Number GCAL LG340810249 Cut Grade IDEAL/EX/EX	191 -Loose Lab Grown Dia
3.04Ct Round Brilliant E VS1 Lab Grown Diamond Cert Number IGI Lg662424955 Cut Grade Ex/Ex/Ex	191 -Loose Lab Grown Dia
0.71 ct Round Brilliant Cut Diamond D SI1	190 -Dia -Loose
2.00 Ct Oval Brilliant Cut Diamond H SI1 GIA Certifcate and laser inscription 5406232622	190 -Dia -Loose
Roberto Coin 18 Karat Yellow Gold Navarra Bangle Bracelet With 60=0.72Tw Round Diamonds	170 -Dia Brac
Roberto Coin Rose Gold 18 Karat Love In Verona Bracelet With 1.87Tw Round Black & White Diamonds	170 -Dia Brac
Marco Bicego 18K Yellow Gold Marrakech Twisted Coil Bracelet With Diamonds Diamonds total carat weight 0.15ctw Length 7 Inch	170 -Dia Brac
Forevermark 18 Karat Yellow Gold Cuff Bracelet With One 0.20Ct Forevermark Round Diamond And One 0.14Ct Forevermark Round Diamond FM 7816757 / 2261013	170 -Dia Brac
14 Karat Yellow Gold Tennis Bracelet With 5=0.53Tw Baguette Stations Diamonds And 74=1.78Tw Round Diamonds Length/Size 6.75"	170 -Dia Brac
Hearts On Fire 18 Karat White Gold Vela Twisted Diamond Pendant 1.17 Total Diamond Weight 18 Inch	165 -Dia Neck
18/14 Karat White Gold One diamond Necklace With One 0.19Ct Round Diamond Length 16/18"	165 -Dia Neck
18 Karat Rose Gold Circle Pendant With 109=0.42Tw Round Diamonds And 20=0.72Tw Round Brown Diamonds on 18 inch yellow gold box chain	160 -Dia Pd
14 Karat Yellow Gold Polished Scattered Diamonds Medium HoopEarrings With 0.49Tw Round H/I Si1-2 Diamonds	150 -Dia Ear
Verragio 14 Karat White Gold Renaissance Semi-Mount Engagement Ring With Round Diamonds 0.59 Total Diamond Weight	140 -Dia Semi-Mt Ring
Forevermark 18 Karat White Gold Ring With One 0.31Ct Forevermark Pear I Vs2 Diamond And 104=0.61Tw Non Forevermark Round Diamonds FM 2267032	100 -Dia Bridal Ring (complete)

Problem: We have a freeform text field that we can naively assume...

- Contains all relevant information in one place
- Is written differently from client to client
- Often includes shorthand, abbreviations, and misspellings
- Often overly verbose or not verbose enough
- May lack specific information that can be instead inferred using context clues

Objective

- Create a parsing solution that operates quickly
- Solution must handle a wide range of written behaviors
- Output must be consistent and predictable
- Can be customized quickly case by case, as needed
- Easy to teach other developers how to use

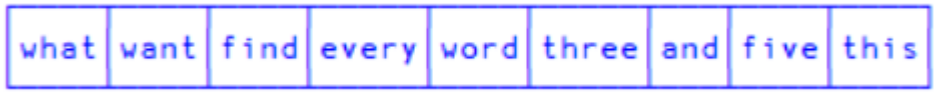
BIG

Immediate concerns

Who has ever worked with freeform data?
Who is having flashbacks already?

We are working with millions of product descriptions written in this manner on any given day

str←'what if we want to find every word between three and five letters in this sentence'
'\b\p{L}{3,5}\b'⎕S'\l0'⊢str



BIG

What sets BIG apart

- Data parsing services require customer to do a lot of the legwork
  - Registering each product as separate item
  - Inputting many specific details into separate fields, for each product
- Jewelry is specific enough that we can do the work for them
- Useful internally for future data science projects/plans

Poor use-case for AI

- Equivalent of trying to hammer a nail with a sledgehammer
  - Outsized tool for the job
- Unpredictable results
- Fails when given complex instructions such as parsing many different items from text
- Difficult to customize output client-to-client

BIG

General strategy to parse anything

- Build data dictionaries per item to parse for a given field
  - Column 1: Type
  - Column 2: Variant alias
  - Column 3: Output format (regex)
  - Column 4: Match regex
- Pre/post processing to make stuff look pretty
- Two lines of code

```
regex_dictionary←⎕CSV 'dictionary.csv'
matches←((regex_dictionary[;4])⎕S(regex_dictionary[;3]))¨description
```

Regex dictionary example: Materials

chrome	cobalt	cobalt chrome	(cobalt chrome|chrome cobalt)
cobalt	cobalt	cobalt	cobalt
copper	copper	copper	copper
enamel	enamel	enamel	enamel
fossil	fossil	fossil	fossil
glass	glass	glass	glass
gold	white	white gold	(?<![."0-9a-z])[12][0-9][^a-z0-9]*?(k(arat)?)?[^a-z0-9]*w(h(i(te?)?)?)?[^a-z0-9]*(g(old)?)?|(?<!(y(e(l(l(ow?)?)?)?)?|r(o(s
gold	yellow	yellow gold	(?<![."0-9a-z])[12][0-9][^a-z0-9]*?(k(arat)?)?[^a-z0-9]*y(e(l(l(ow?)?)?)?)?[^a-z0-9]*(g(old)?)?|(?<!(w(h(i(te?)?)?)?|r(o(s
gold	rose	rose gold	(?<![."0-9a-z])[12][0-9][^a-z0-9]*?(k(arat)?)?[^a-z0-9]*r(o(se?)?)?[^a-z0-9]*(g(old)?)?|(?<!(w(h(i(te?)?)?)?|y(e(l(l(ow?)'
gold	green	green gold	(?<![."0-9a-z])[12][0-9][^a-z0-9]*?(k(arat)?)?[^a-z0-9]*gre?e?n[^a-z0-9]*(g(old)?)?|(?<!(w(h(i(te?)?)?)?|y(e(l(l(ow?)?)?
gold	twotone	two-tone gold	(?<![."0-9a-z])[12][0-9][^a-z0-9]*?(k(arat)?)?[^a-z0-9]*t(wo[^a-z0-9]*)?t(one)?[^a-z0-9]*(g(old)?)?|(?<!(w(h(i(te?)?)?)?
gold	gold	gold	(?<![."0-9a-z])[12][0-9][^a-z0-9]*?(k(arat)?)?[^a-z0-9]*(?<!(w(h(i(te?)?)?)?|y(e(l(l(ow?)?)?)?)?|r(o(se?)?)?|gre?e?n|t(w
leather	leather	leather	leather
meteorite	meteorite	meteorite	meteorite
mop	mop	\1 mother of pearl	(ivory?| ?)(mother\s*-*\s*of\s*-*\s*pearl|m\.o\.p\.?|mop|nacre)

The association problem

- What if we have a single product description that
  - Contains multiple gemstones
  - Each gemstone has specific associated attributes
  - Attributes associated with gemstones have varying conditions
    - One carat max per gemstone group
    - One count per gemstone group
    - Some gemstones might not be gemstones (emerald)
    - Sometimes gemstones not explicitly mentioned

- How do we associate certain qualities to certain gemstones?

BIG

Building a stone string – Ex 1

18 Karat two-tone ring set with a 0.51ct Pear Diamond (14) Round Diamonds at 0.34cts F VS (11) Round Diamonds at 0.28cts G/H VS

Building a stone string – Ex 1

18 Karat two-tone ring set with a 0.51ct Pear Diamond (14) Round Diamonds at 0.34cts F VS (11) Round Diamonds at 0.28cts G/H VS

Stone

Building a stone string – Ex 1

18 Karat two-tone ring set with a 0.51ct Pear Diamond (14) Round Diamonds at 0.34cts F VS (11) Round Diamonds at 0.28cts G/H VS

Stone

Shape

Building a stone string – Ex 1

18 Karat two-tone ring set with a 0.51ct Pear Diamond (14) Round Diamonds at 0.34cts F VS (11) Round Diamonds at 0.28cts G/H VS

- Stone
- Shape
- Carat

Building a stone string – Ex 1

18 Karat two-tone ring set with a 0.51ct Pear Diamond (14) Round Diamonds at 0.34cts F VS (11) Round Diamonds at 0.28cts G/H VS

- Stone
- Shape
- Carat
- Color + Clarity

Building a stone string – Ex 1

18 Karat two-tone ring set with a 0.51ct Pear Diamond (14) Round Diamonds at 0.34cts F VS (11) Round Diamonds at 0.28cts G/H VS

- Stone
- Shape
- Carat
- Color + Clarity
- Wildcard

Building a stone string – Ex 1

18 Karat two-tone ring set with a [0.51ct Pear Diamond] (14) [Round Diamonds at 0.34cts F VS] (11) [Round Diamonds at 0.28cts G/H VS]

- Stone
- Shape
- Carat
- Color + Clarity
- Wildcard
- Stone strings

0.51ct Pear Diamond

Round Diamonds at 0.34cts F VS

Round Diamonds at 0.28ctw G/H VS

Building a superpattern

- Previous example was ideal
  - All parsed items close together
  - Everything properly spelled
- Parse each item separately
  - Stone size (1/2/3D in mm)
  - Stone qualities (‘pink’ in pink diamond’
  - Stone shape (round, pear, etc)
  - Stone cut
  - Stone color/clarity
  - Carat weights
- Turn each group of matches into regex ‘capture group’
- Prefix modifier (with carat)
  - PC
- Prefix modifier (without carat)
  - P
- Stone type
  - X
- Suffix modifier (with carat)
  - SC
- Suffix modifier (without carat)
  - S
- Wildcard and Optional
  - + and ?

BIG

Building a superpattern

- Prefix modifier (with carat)
  - PC
- Prefix modifier (without carat)
  - P
- Stone type
  - X
- Suffix modifier (with carat)
  - SC
- Suffix modifier (without carat)
  - S
- Wildcard and Optional
  - + and ?

In order of priority:

1) PC + X + S
2) PC + X
3) P + X + SC
4) X + SC
5) P? + X + S?

Revisiting Ex 1

18 Karat two-tone ring set with a 0.51ct Pear Diamond (14) Round Diamonds at 0.34cts F VS (11) Round Diamonds at 0.28cts G/H VS

- Stone
- Shape
- Carat
- Color + Clarity
- Wildcard
- Stone strings

PC X
0.51ct Pear Diamond

P X + SC
Round Diamonds at 0.34cts F VS

P X + SC
Round Diamonds at 0.28ctw G/H VS

Messy descriptions – Ex 2

MQ.50C T 11=225CTW MLEE LG PSAP.16 CT 14KW

Messy descriptions – Ex 2

MQ.50C T	11=225CTW MLEE LG	PSAP.16 CT	14KW

P X/C
MQ.50C T
X/C S
11=225CTW MLEE LG
P X SC
PSAP.16 CT

- Stone
- Shape
- Carat
- Color + Clarity
- Wildcard
- Stone strings

How it works

- Each step, we use regex dictionaries to reduce problem scope
  - Find all matches
  - Create regular expression from matches
  - Used reduced scope capture groups to build super patterns
  - Feed superpatterns back in, get new matches
- Superpattern matches aren’t the final step
  - Create new ordered regex capture group using ALL superpatterns
  - Get final matches, using single sweep
- Last step necessary to prevent overlapping stone string

BIG

How it works

- Scope of the problem is massive and leads to impossibly complex regex

- Dynamically build regular expressions piece by piece
  - Cons: Many more regex operations
  - Pros: Code is easy to diagnose when running into issues

BIG

Down the rabbit hole

- What if nucleation point can be multiple categories?
  - Example: ‘Diamond’ may be expressed as ‘G/H VS1-VS2’
  - Decode dictionary may be different description to description
- What if nucleation point can be abbreviated?
  - Example: ‘RD SA DI 0.25CTW’
  - Potentially captured in wildcard, blurring boundaries of stone strings

BIG

Down the rabbit hole

- What if description written with unnatural English?
  - Example ‘0.25 CTW DIA PINK
- What if we capture multiple stones but there is only one stone?
  - Example ‘LOOSE DIAMOND RING WITH 0.25CTW DIA’

Built in Dyalog

No :For loops outside of customization file

- Inner products
  - ,.{}
- Reduction and outer products
  - ,⌿∘.{}
- Dfns to quickly handle :If, very useful for regex construction
  - {×≢ω: ω ⋄ ''}¨
  - Transparent
- Key operator
  - {}⌸

Thank you!

DYNA ’26 – 04/27/2026

Mark Wolfson and Kori Smith