Furever Match

A data product helping kitties everywhere
find their furever home

Alexandra Plassaras

What kind of impact could this really have?

93.8 M

Cats owned in the US (2017)

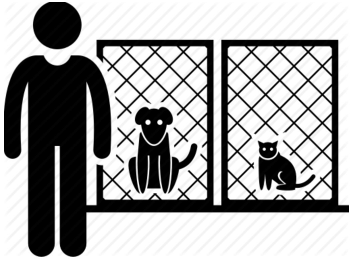
7.6 M

Cats and Dogs re-enter shelters annually

Case Study: National vs. hyperlocal data

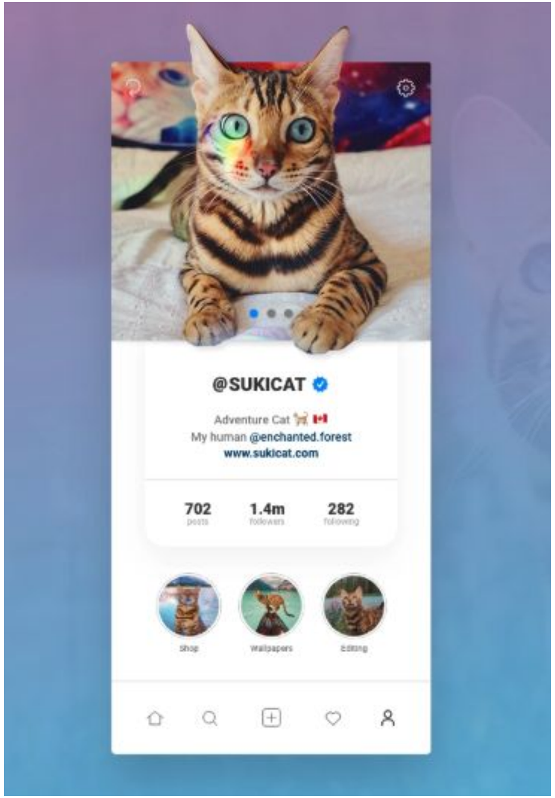
petfinder®

+



King County

=



@SUKICAT

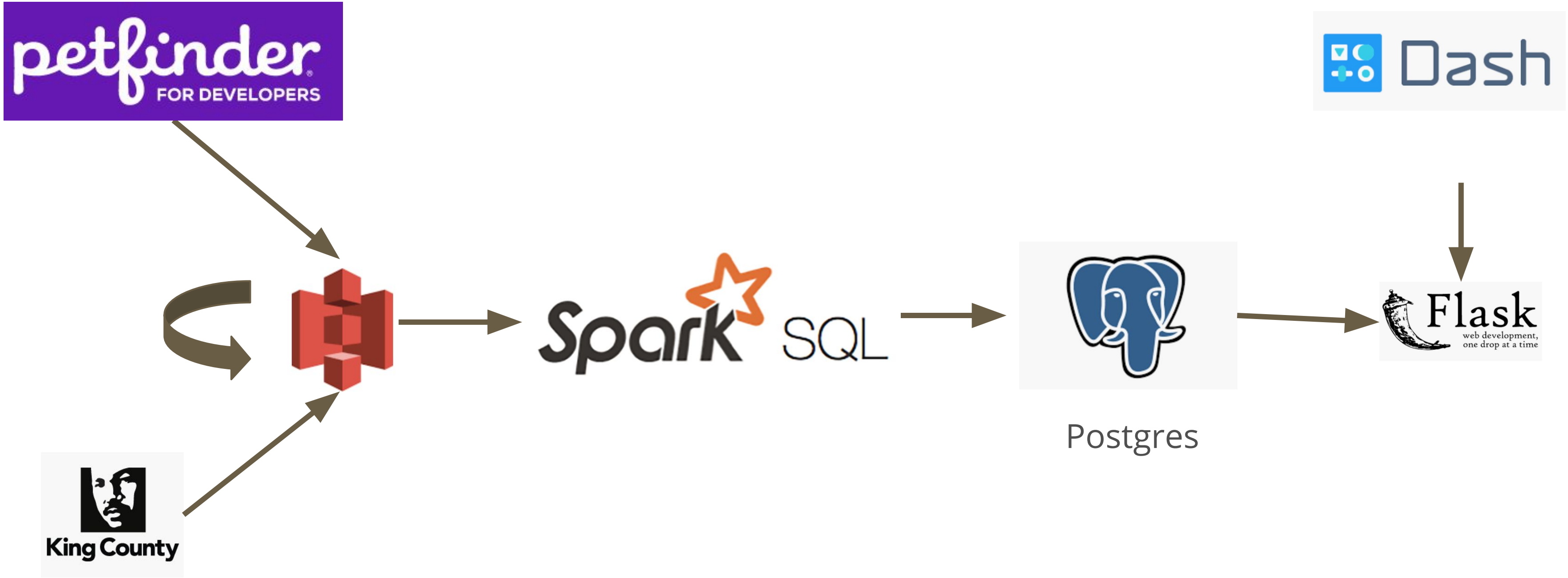
Adventure Cat 🐈 🇨🇦
My human @enchanted.forest
www.sukicat.com

702	1.4m	282
posts	followers	following

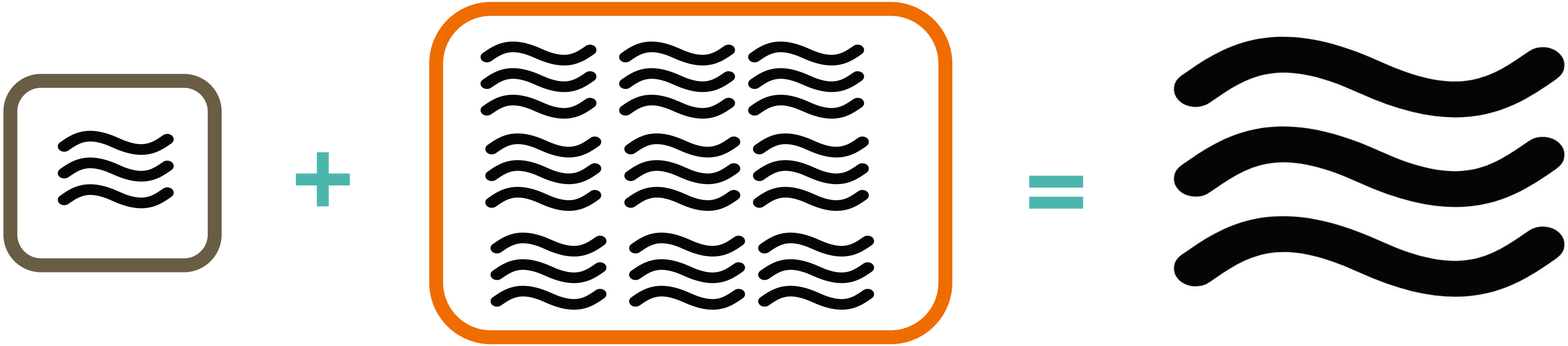
Shop · Wallpapers · Editing

Result: A tool for volunteers, analysts and engineers

Batch Processing Data Pipeline



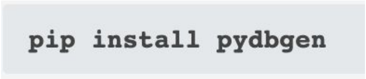
Challenge: Acquire data



Raw data - Comprised of a subset of states with real cats

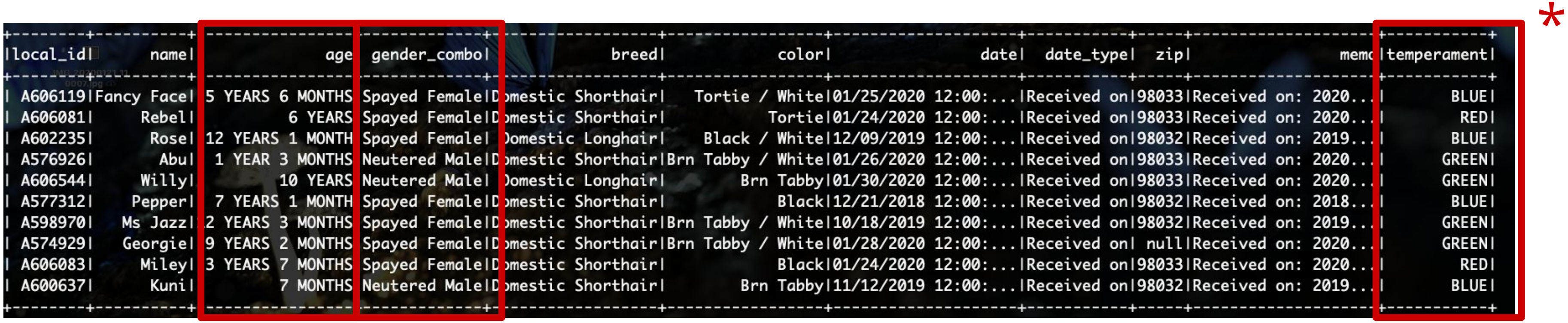
Synthetic data - Randomly generated data from all 50 states, top 100 popular cat names

mockaroo



```
pip install pydbgen
```

Challenge: Data Wrangling



```
+--------+----------+----------------+-------------+------------------+-----------------+--------------------+-----------+-----+--------------------+-----------+
|local_id|      name|             age| gender_combo|             breed|            color|                date|  date_type|  zip|                memo|temperament|
+--------+----------+----------------+-------------+------------------+-----------------+--------------------+-----------+-----+--------------------+-----------+
| A606119|Fancy Face|5 YEARS 6 MONTHS|Spayed Female|Domestic Shorthair|   Tortie / White|01/25/2020 12:00:...|Received on|98033|Received on: 2020...|       BLUE|
| A606081|     Rebel|         6 YEARS|Spayed Female|Domestic Shorthair|           Tortie|01/24/2020 12:00:...|Received on|98033|Received on: 2020...|        RED|
| A602235|      Rose|12 YEARS 1 MONTH|Spayed Female| Domestic Longhair|    Black / White|12/09/2019 12:00:...|Received on|98032|Received on: 2019...|       BLUE|
| A576926|       Abu| 1 YEAR 3 MONTHS|Neutered Male|Domestic Shorthair|Brn Tabby / White|01/26/2020 12:00:...|Received on|98033|Received on: 2020...|      GREEN|
| A606544|     Willy|        10 YEARS|Neutered Male| Domestic Longhair|        Brn Tabby|01/30/2020 12:00:...|Received on|98033|Received on: 2020...|      GREEN|
| A577312|    Pepper| 7 YEARS 1 MONTH|Spayed Female|Domestic Shorthair|            Black|12/21/2018 12:00:...|Received on|98032|Received on: 2018...|       BLUE|
| A598970|   Ms Jazz|2 YEARS 3 MONTHS|Spayed Female|Domestic Shorthair|Brn Tabby / White|10/18/2019 12:00:...|Received on|98032|Received on: 2019...|      GREEN|
| A574929|   Georgie|9 YEARS 2 MONTHS|Spayed Female|Domestic Shorthair|Brn Tabby / White|01/28/2020 12:00:...|Received on| null|Received on: 2020...|      GREEN|
| A606083|     Miley|3 YEARS 7 MONTHS|Spayed Female|Domestic Shorthair|            Black|01/24/2020 12:00:...|Received on|98033|Received on: 2020...|        RED|
| A600637|      Kuni|        7 MONTHS|Neutered Male|Domestic Shorthair|        Brn Tabby|11/12/2019 12:00:...|Received on|98032|Received on: 2019...|       BLUE|
+--------+----------+----------------+-------------+------------------+-----------------+--------------------+-----------+-----+--------------------+-----------+
```

*

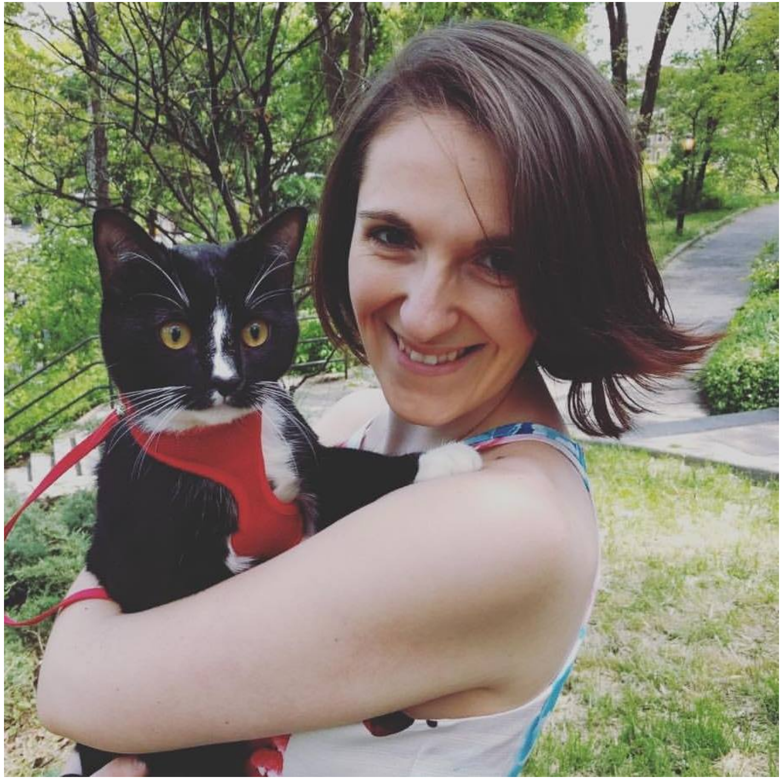
- Age is a string with both years and months
- Gender is gender plus spayed/neutered
- Temperament is on a scale compared to good with children, cats and dogs

Alexandra Plassaras

M.S. in Quantitative Methodologies for Social Science

COLUMBIA UNIVERSITY

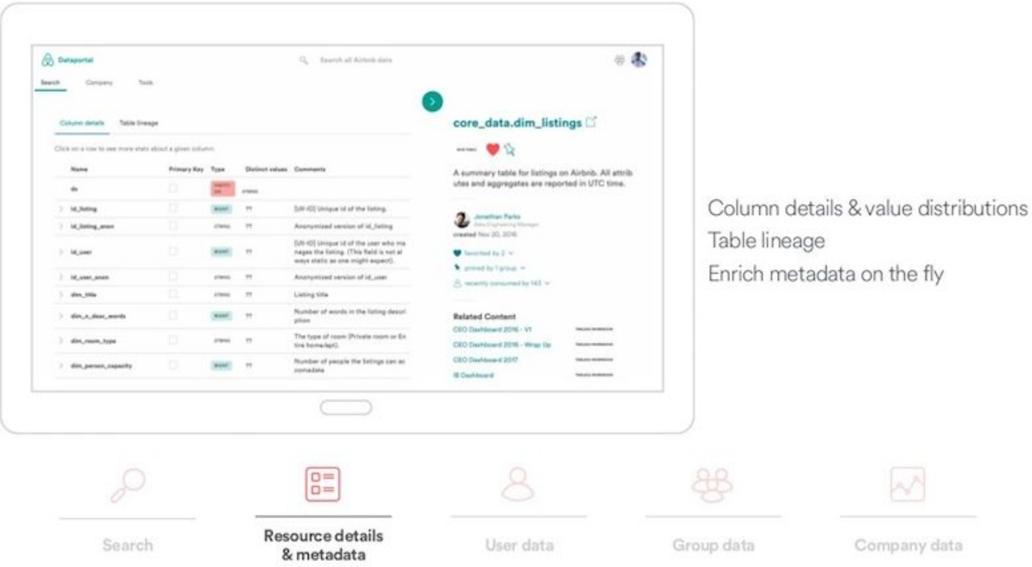
Udemy



aplassaras@gmail.com

What’s next for furever match?

- Build out backend
  - Airflow to automate data ingestion and prioritize data
  - Store historical snapshots
- Bring in the Data Analysts / Scientists
  - NLP for textual data
  - Image processing against adoptable and lost pets
  - Recommendation algorithms for owner + cat



Click on a row to see more stats about a given column.

Name	Primary Key	Type	Distinct values	Comments
ds	□	[illegible]	STRING	
id_listing	□	BIGINT	??	[UII-ID] Unique id of the listing.
id_listing_anon	□	STRING	??	Anonymized version of id_listing
id_user	□	BIGINT	??	[UII-ID] Unique id of the user who manages the listing. (This field is not always static as one might expect).
id_user_anon	□	STRING	??	Anonymized version of id_user
dim_title	□	STRING	??	Listing title
dim_n_desc_words	□	BIGINT	??	Number of words in the listing description
dim_room_type	□	STRING	??	The type of room (Private room or Entire home/apt).
dim_person_capacity	□	BIGINT	??	Number of people the listings can accomadate

core_data.dim_listings

A summary table for listings on Airbnb. All attributes and aggregates are reported in UTC time.

Jonathan Parks
Data Engineering Manager
created Nov 20, 2016

favorited by 2
pinned by 1 group
recently consumed by 143

Related Content
CEO Dashboard 2016 - V1
CEO Dashboard 2016 - Wrap Up
CEO Dashboard 2017
IB Dashboard

Column details & value distributions
Table lineage
Enrich metadata on the fly
Search
Resource details & metadata
User data
Group data
Company data

Why I used S3

- Easily integrate with current pipeline and with other AWS services
- Durability of 99.999999999% of objects per year
- Data on S3 persists compared to HDFS which doesn’t persist once an instance is stopped
- With S3 you only pay for the storage that you actually need plus S3 compresses files

Why I used Spark?

- Faster than Hadoop
    - Spark tries to do as many calculations as possible in memory, which avoids moving data back and forth across a cluster
- Apache Spark is known for batch-processing big data
- It is open source analytics platform for large-scale processing of huge datasets, large online community
- It has resilient distributed datasets (RDDs), and the in-memory data structure allows Spark to perform functional programming.
- It uses a DAG scheduler along with physical execution engine and the query optimizer.
- It contains a stack of libraries including Spark SQL

Why I used RDS PostgreSQL (compared to Redshift)

- Scaling:
  - Takes only a few minutes for RDS (reconfiguring virtual instances)
- Storage Capacity:
  - I don’t need a storage capacity of up to 2 PB
- Data Replication:
  - Don’t need to copy complete data to S3 and then copy (Redshift).
  - Depends on the underlying database I use in RDS

- Pricing:
  - Limited budget, RDS already included in current AWS services ($0.017 vs $0.25)
- Performance:
  - RDS has better performance for queries that don’t tests its limits (millions of rows)
    - Given my budget for AWS and $0 budget for APIs Isn’t an issue for now
- What I could do in the future:
  - Once my data is sufficiently large I could add Redshift to my robust analytical pipeline

Challenge: Acquire data

Raw data

Duplicate data

Synthetic data

JSON

Animal info

Animal Description

Organization Info

Temperament

Medical Info

Adoption Status