Extraction, Transformation,

and Load Technical Report

Animal Adoption Decision Maker

1. Introduction 3

1.1 Summary 3

1.2 Scope 3

1.3 Technologies and resource contributions 3

1.4 Definitions, Acronyms and Abbreviations 3

2. ETL Details 4

2.1 Data Import/Extract Sources and Method 4

2.2 Data Acquisition 4

2.3 Data Transform 4

2.4 Data Integrity 4

2.5 Data Refresh Frequency 4

2.6 Data Security 4

2.7 Data Loading and Availability 5

3. Data Quality 6

**1. INTRODUCTION**

**1.1 Summary**

Objective: The objective is to gather pet data using API’s from multiple sources using Python and MongoDB to assist with pet adoption decision making.

Problem Statement: Animal adoption is a big decision for a family. There are many adoptable animals in the world and people can find animal data online. We would like to provide available animal data with Petfinder API, Kaggle.com, ASPCA.com, TheSimpleDollar.com and BLS.com. People can find a match based on their preference such as breed, age, gender, color, temperament, location and visualize current available animals by category. We will analyze the pet industry and provide related cost or expenditure after adoption to help them make a decision.

Expected Outcome of ETL: People make a decision on pet adoption based on initial and lifetime cost. The outcome of the results will provide analysis on pet type/ breed, cost of living and capital cost. By allowing users to search for the perfect pet, the analysis provide will allow individuals to locate pets in other cities and states across the U.S. In addition, we will be able to perform analysis on which pets are not being adopted in a geographical region.

**1.2 Scope**

We could get more than 10,000 available adoptable animal data from nearly 14,000 shelters and rescue groups across North America from Petfinder API. We also collected initial and lifetime pet cost data by animal, size and breed from various sources.

**1.3 Technologies and resource contributions**

Helen Tan – Identified data sources, collected initial and lifetime pet cost data, processed data with Python and loaded on MongoDB.

Jin Kim – Identified data sources, extracted animal data from Petfinder API using python and

transformed data to json and loaded on MongoDB.

Natalie Maize – Converted pdf files csv, data cleansing and collaborated with team members to complete the technical paper.

**1.4 Definitions, Acronyms and Abbreviations**

Acronyms:

* ETL: Extract, Transform, Load
* API: Application Programming Interface
* State: All 50 US states will be abbreviated (ex. GA – Georgia, NY – New York, MS – Mississippi, FL – Florida)
* Pet size: Size of pets will be abbreviated (ex. SM – Small, MED – Medium, LG – Large)

**2. ETL DETAILS**

**2.1 Data Import/Extract Sources and Method**

PetFinder.com: URL - <https://www.petfinder.com/developers/v2/docs/>

* We used API to extract available animal data. To obtain authentication, the PetFinder API key and Secret Key was required. For additional security, there was an access token for the API requests.



Key parameters are following:

category = animals | types | organizations

parameters = type | breed | size | gender | age | color | coat | status | name |

organization | location | distance | sort | page | limit

Detailed parameters are written in Petfinder API documentation.



Kaggle.com: URL - <https://www.kaggle.com/c/petfinder-adoption-prediction/data>

* Source was eliminated due to the lack of data integrity

BLS.com: URL - <https://www.bls.gov/opub/ted/2017/households-spent-an-average-of-528-dollars-on-pets-in-2015.htm>

Thesimpledollar.com: URL <https://www.thesimpledollar.com/pet-cost-calculator/>

* Source to obtain capital expenditure data for pets by household and lifetime pet costs.

ASPCA.org: URL - <https://www.aspca.org/sites/default/files/pet_care_costs.pdf>

* Source to obtain capital expenditure data for pets by household and lifetime pet costs.

**2.2 Data Acquisition**

Available animal data is dynamic and needs update real time. We can pull specific adoptable animal data adding detailed parameters from Petfinder API if needed. Pet costs data is static but can be updated yearly based on current market price.

**2.3 Data Transform**

Below is a list of data elements that were transformed:

* Convert Linux to Python
* Populated Null values when pet photos did not exist and not available
* City, State, Zip
* Format column headers
* Pdf to csv – changed the object to floats

**2.4 Data Integrity**

Overall, the datasets were not lacking data integrity. Although, the datasets provided additional information that will not be useful for analysis. The response from API call has unnecessary data and some essential data is difficult to access because they are provided by multiple nested object. Therefore, we make a new dictionary object and clean up unnecessary one-by-one data and store it on MongoDB as well as a json file.

**2.5 Data Refresh Frequency**

Animal adoptions happen every day. To get the most recent update available animal data real time, we recommend refresh data on a daily basis.

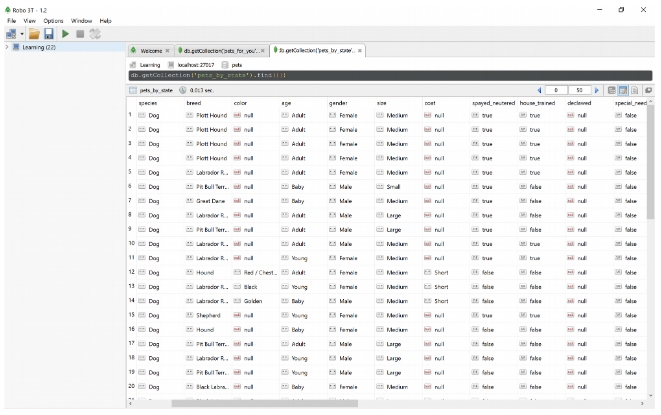
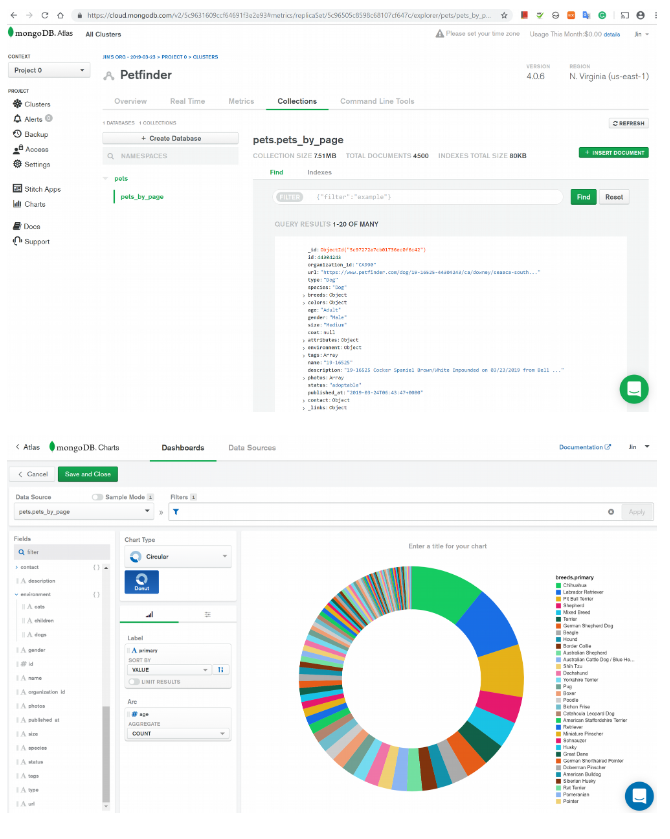
**2.6 Data Security**

Our data set only have public information and does not include sensitive information currently. We plan to implement elevated data security level after we add clients (animal adopters) collection on our MongoDB. We will mask customers name partially and encrypt privacy data such as address or phone numbers. Also, we will have backup collections of adopted animals separately.

2.7 **Data Loading and Availability**

We loaded our data on mongoDB. Users can access data with mongoDB compass or alternatives such as Robo 3T, Studio 3T on local server. Also we connected and loaded the data on cloud mongoDB and it can be accessed with the URL below only permitted users with ID and password and IP Whitelist.

https://cloud.mongodb.com/v2/5c9631609ccf64691f3e2e93#clusters

**3. DATA QUALITY**

We aim to potential animal adopters can find their perfect match from our dataset. Clients can access more than 100,000 adoptable animal data from nearly 14,000 shelters and rescue groups across North America by type, breed, age, color, size, location and more. We tested sample animal data from Atlanta, GA and confirmed that they are also available from local shelter websites.