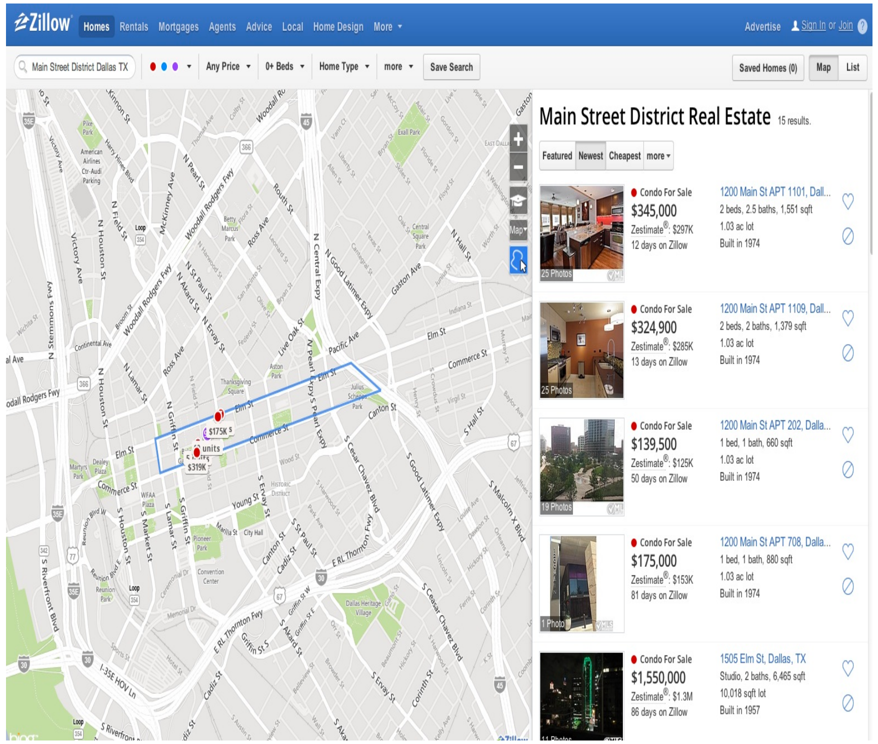
Zillow End-to-End Data Pipeline

BY JERRY KHONG

Introduction

- Zillow is a real estate marketplace platform
- Buy, sell, rent
- 160 million active monthly users [1]

Zillow®



Main Street District Real Estate 15 results.

Condo For Sale
$345,000
Zestimate®: $297K
12 days on Zillow
1200 Main St APT 1101, Dall...
2 beds, 2.5 baths, 1,551 sqft
1.03 ac lot
Built in 1974

Condo For Sale
$324,900
Zestimate®: $285K
13 days on Zillow
1200 Main St APT 1109, Dall...
2 beds, 2 baths, 1,379 sqft
1.03 ac lot
Built in 1974

Condo For Sale
$139,500
Zestimate®: $125K
50 days on Zillow
1200 Main St APT 202, Dalla...
1 bed, 1 bath, 660 sqft
1.03 ac lot
Built in 1974

Condo For Sale
$175,000
Zestimate®: $153K
81 days on Zillow
1200 Main St APT 708, Dalla...
1 bed, 1 bath, 880 sqft
1.03 ac lot
Built in 1974

Condo For Sale
$1,550,000
Zestimate®: $1.3M
86 days on Zillow
1505 Elm St, Dallas, TX
Studio, 2 baths, 6,465 sqft
10,018 sqft lot
Built in 1957

[1]: HTTP://EXPANDEDRAMBLINGS.COM/INDEX.PHP/ZILLOW-STATISTICS/

galvanizeU
powered by University of New Haven

Problems in Real Estate

- U.S. residential real estate market is one of the least liquid markets
- Real estate market is volatile and hard to predict
- Real estate transactions are stressful
- Zillow's estimated house valuations are bad

What is the true value of a home?

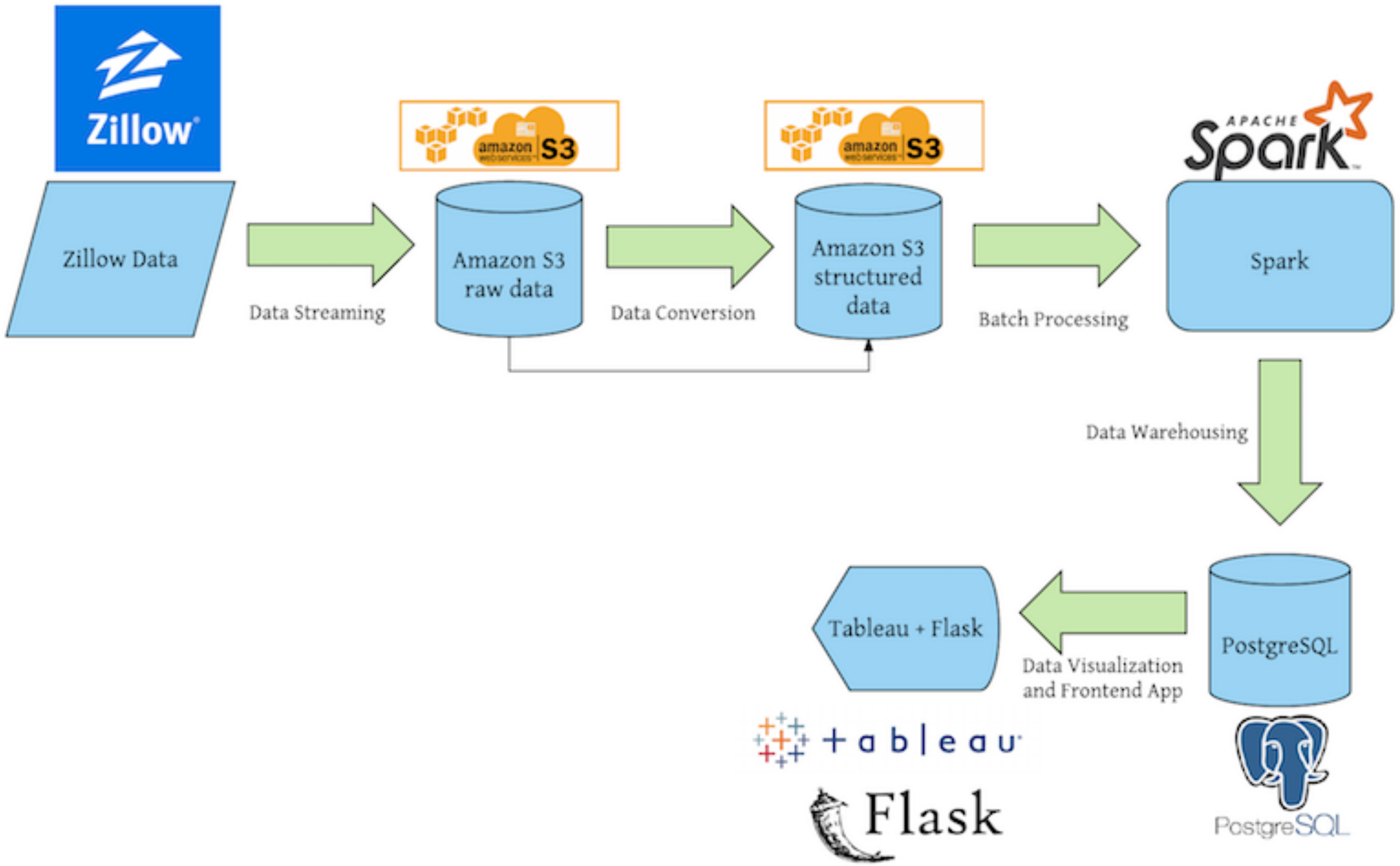
Solution

- Assess which areas will be good real estate investments based off historical and current market data.

- Provide users with interactive map to see which zip codes are currently under-valued and over-valued.

- Future product: scale out to display individual listings

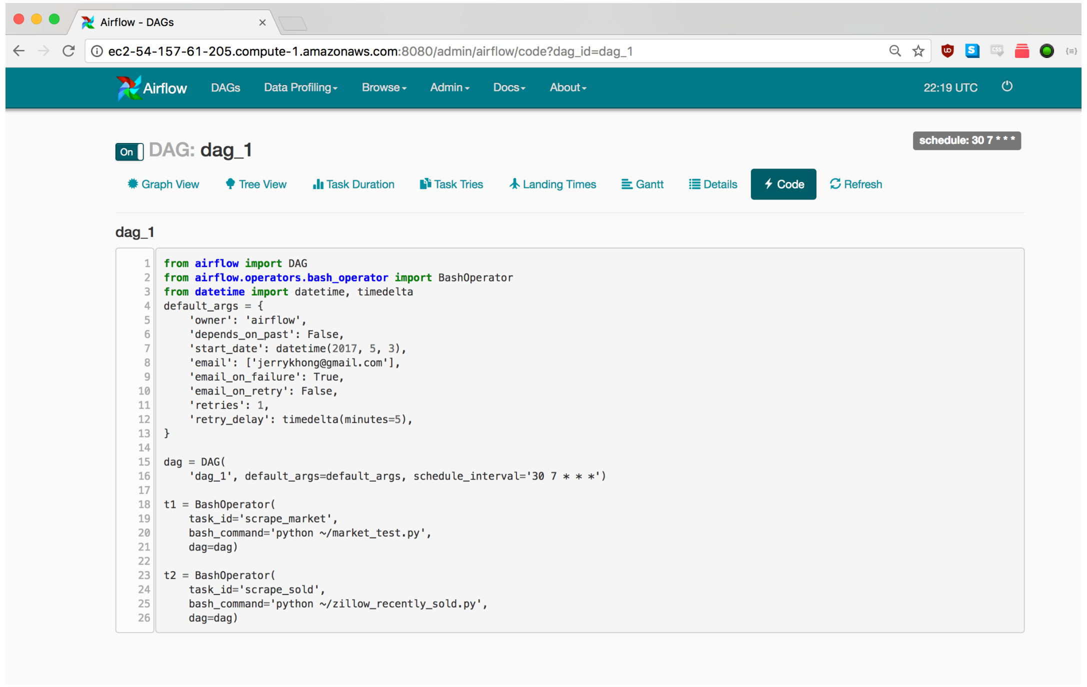
Data Pipeline Architecture



Stream

- EC2 instance bootstrapped with Selenium and Chrome

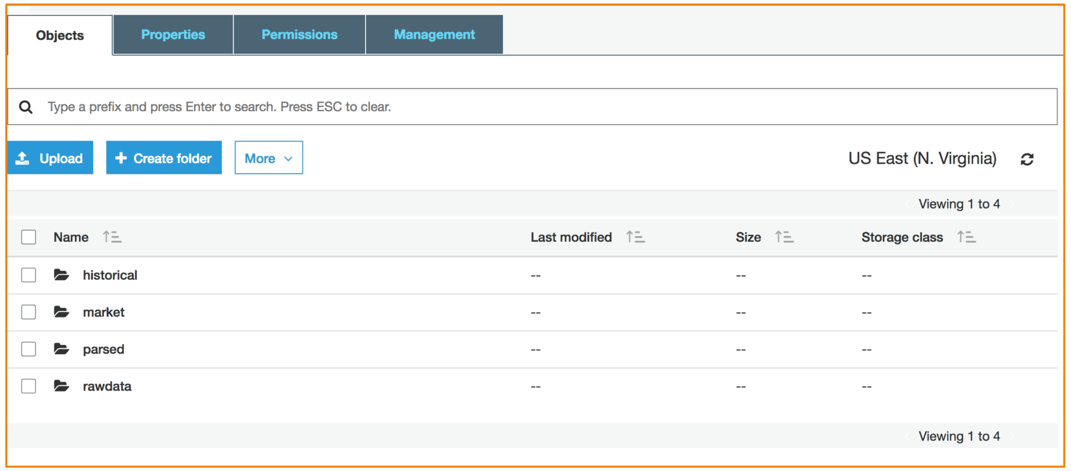
- Setup Airflow DAG (daily)
  I. Scrape all houses on the market
  II. Scrape all recently sold houses



```python
from airflow import DAG
from airflow.operators.bash_operator import BashOperator
from datetime import datetime, timedelta
default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': datetime(2017, 5, 3),
    'email': ['jerrykhong@gmail.com'],
    'email_on_failure': True,
    'email_on_retry': False,
    'retries': 1,
    'retry_delay': timedelta(minutes=5),
}

dag = DAG(
    'dag_1', default_args=default_args, schedule_interval='30 7 * * *')

t1 = BashOperator(
    task_id='scrape_market',
    bash_command='python ~/market_test.py',
    dag=dag)

t2 = BashOperator(
    task_id='scrape_sold',
    bash_command='python ~/zillow_recently_sold.py',
    dag=dag)
```

Store

- Store raw HTML data in S3
- Parse raw HTML data into CSV and store in S3

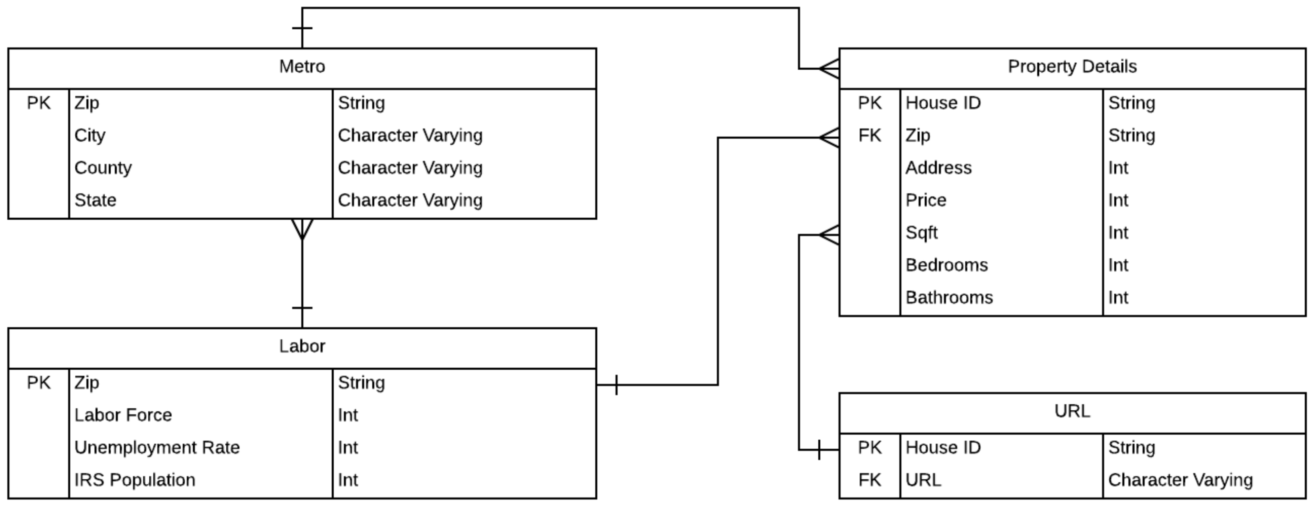


Name	Last modified	Size	Storage class
historical	--	--	--
market	--	--	--
parsed	--	--	--
rawdata	--	--	--

Viewing 1 to 4

Structure

- Use Apache Spark to push parsed data in Postgres DB (3NF)
- Use Apache Spark to push historical data in Postgres DB (3NF)



Metro

PK	Zip	String
	City	Character Varying
	County	Character Varying
	State	Character Varying

Property Details

PK	House ID	String
FK	Zip	String
	Address	Int
	Price	Int
	Sqft	Int
	Bedrooms	Int
	Bathrooms	Int

Labor

PK	Zip	String
	Labor Force	Int
	Unemployment Rate	Int
	IRS Population	Int

URL

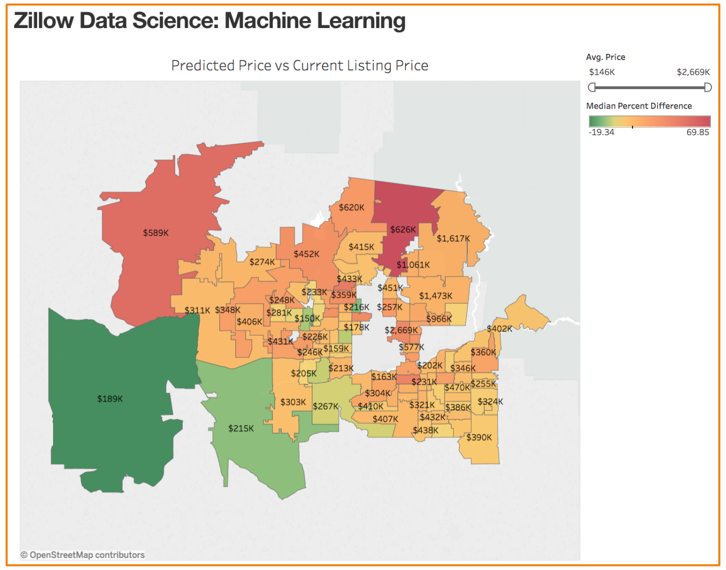
PK	House ID	String
FK	URL	Character Varying

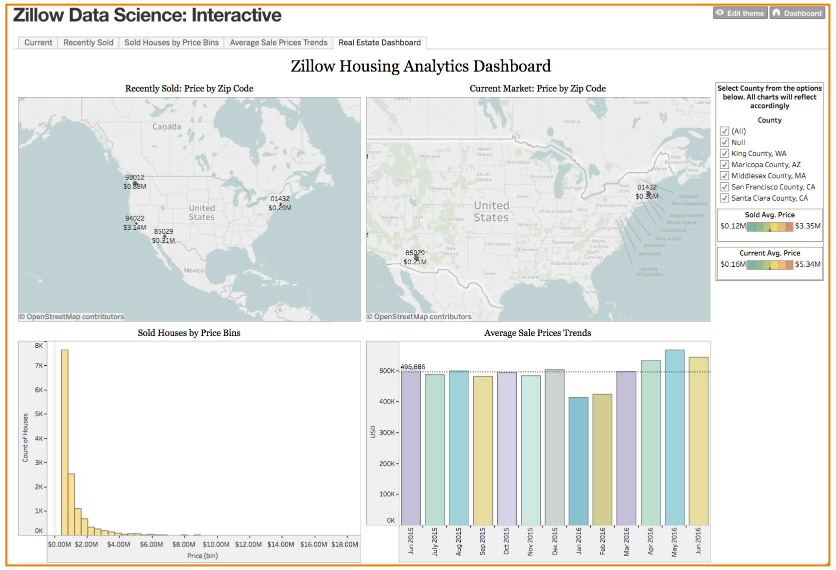
Synthesize

- SQL queries for machine learning → export as CSV file
- XGBoost Regression on individual listings to predict price
- Connect Postgres DB to Tableau
- SQL queries in Tableau to filter median house prices

Show

- Create data visualizations in Tableau
- Embed interactive dashboard on website


Zillow Data Science: Machine Learning
Predicted Price vs Current Listing Price
Avg. Price
$146K
$2,669K
Median Percent Difference
-19.34
69.85
© OpenStreetMap contributors


Zillow Data Science: Interactive
Current
Recently Sold
Sold Houses by Price Bins
Average Sale Prices Trends
Real Estate Dashboard
Zillow Housing Analytics Dashboard
Recently Sold: Price by Zip Code
Current Market: Price by Zip Code
Sold Houses by Price Bins
Average Sale Prices Trends
County
King County, WA
Maricopa County, AZ
Middlesex County, MA
San Francisco County, CA
Santa Clara County, CA

8 Properties for Big Data systems

Robustness and Fault Tolerance

All systems belong to AWS, which is highly robust. In addition, all systems integrates with one another. If the scraping goes down, data itself will be safe in S3.

Low latency reads and updates

Since the real estate sales cycle is relatively slow, there is no need for real time updates.

Scalability

A majority of these technologies are fully scalable. S3 and Spark have high scalability. MySQL or SparkSQL might be a better fit than Postgres for scalability. Tableau is also very scalable.

Generalization

This data pipeline architecture is easily extendible for other scraping and can be reused.

8 Properties for Big Data systems (cntd)

Extensibility

All of the systems in place support the addition of new data sources, features, and models.

Ad hoc queries

Postgres and Tableau will be used to do ad hoc queries.

Minimal maintenance

This system is complex, and requires monitoring to make sure everything is running. I attempted to make an e-mail airflow DAG in case of errors.

Debuggability

Data will always be stored in complete form in S3, so bugs can be easily traced out. Ability for data to be restructured and models can be recomputed if something goes wrong.

Challenges and Potential Risks

- Working with HTML vs JSON
- Scraping vs API usage
- Bootstrapping different packages
- Getting banned from Zillow
- Research shows that Zillow's data is not up to date (slow)

Future Goals

- Display individual listings
- Scrape more features
- Incorporate WalkScore and GreatSchools API