Piloting a new data ecosystem with Google Cloud Platform

Expanding access to NYC Planning’s data products

Spencer Simon, Matt Crittenden - ITD Spotlight, August 2021

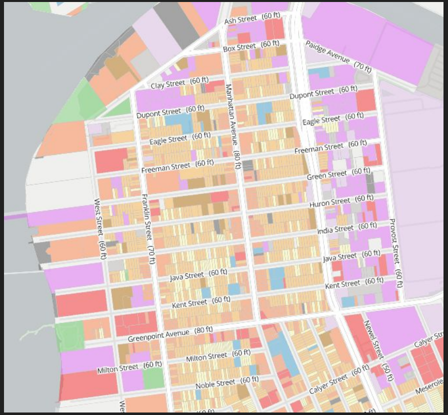
PART ONE

Overview

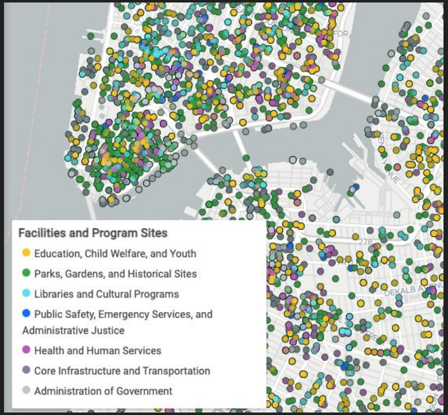
Hello!

Data Engineering’s Data Products

PLUTO



Facilities database


Facilities and Program Sites
● Education, Child Welfare, and Youth
● Parks, Gardens, and Historical Sites
● Libraries and Cultural Programs
● Public Safety, Emergency Services, and Administrative Justice
● Health and Human Services
● Core Infrastructure and Transportation
● Administration of Government
FDR
DEKALB AV

Housing database

Housing Development
● New Building
● Alteration
● Demolition
FDR
FDR
MYR
ALB AVE

and more...

City Owned and Leased Properties,
Known Projects database,
Capital Projects database,
Community District Profiles,
Population Fact Finder,
ZAP,
CEQR ...

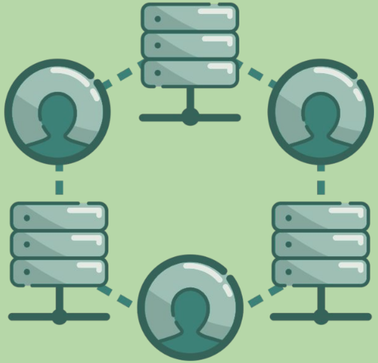
Data Engineering’s Mission

Product

Create and release **high quality public datasets** about NYC.

Operation

Build highly **transparent** and **automated** data pipelines using **open source technologies**.



Ecosystem

Offer more than just data, but also comprehensive **documentation** and **metadata**.

Community

Bring people together, across teams and agencies, to share data and to learn from each other.

A data ecosystem is your environment for working with data

Discovery

The Guide

Access

Bytes,
Open Data,
M Drive,
Carto

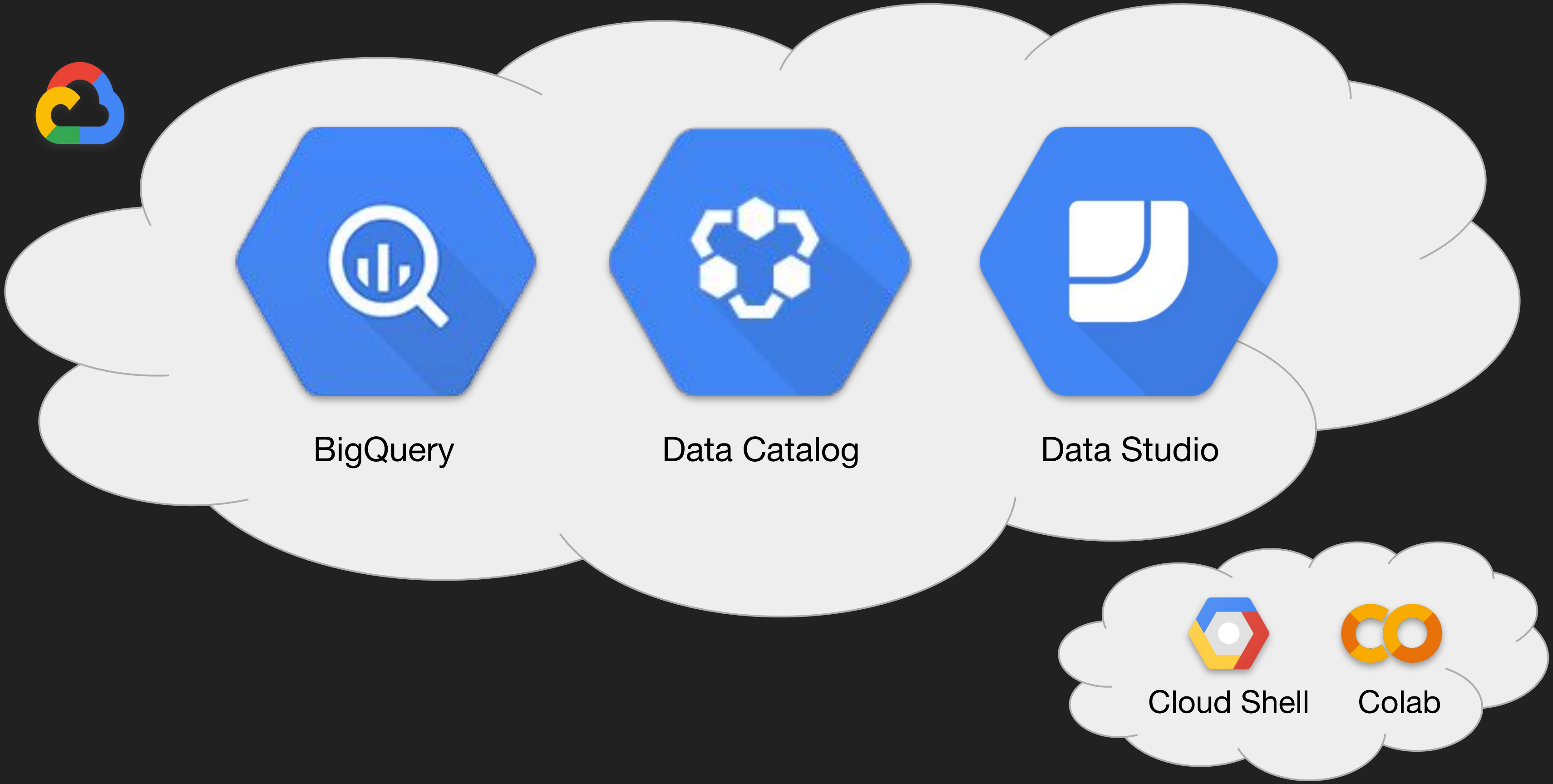
Collaboration

Emails,
Teams,
Meetings

Analysis

Excel,
Tableau,
ArcMap

Enhancing the Data Ecosystem with Google Cloud Platform



BigQuery

Data Catalog

Data Studio

Cloud Shell

Colab

Our Use Case

How have zoning changes impacted the built environment?

Q1: What happens to the number of units in a lot following a zoning district change?

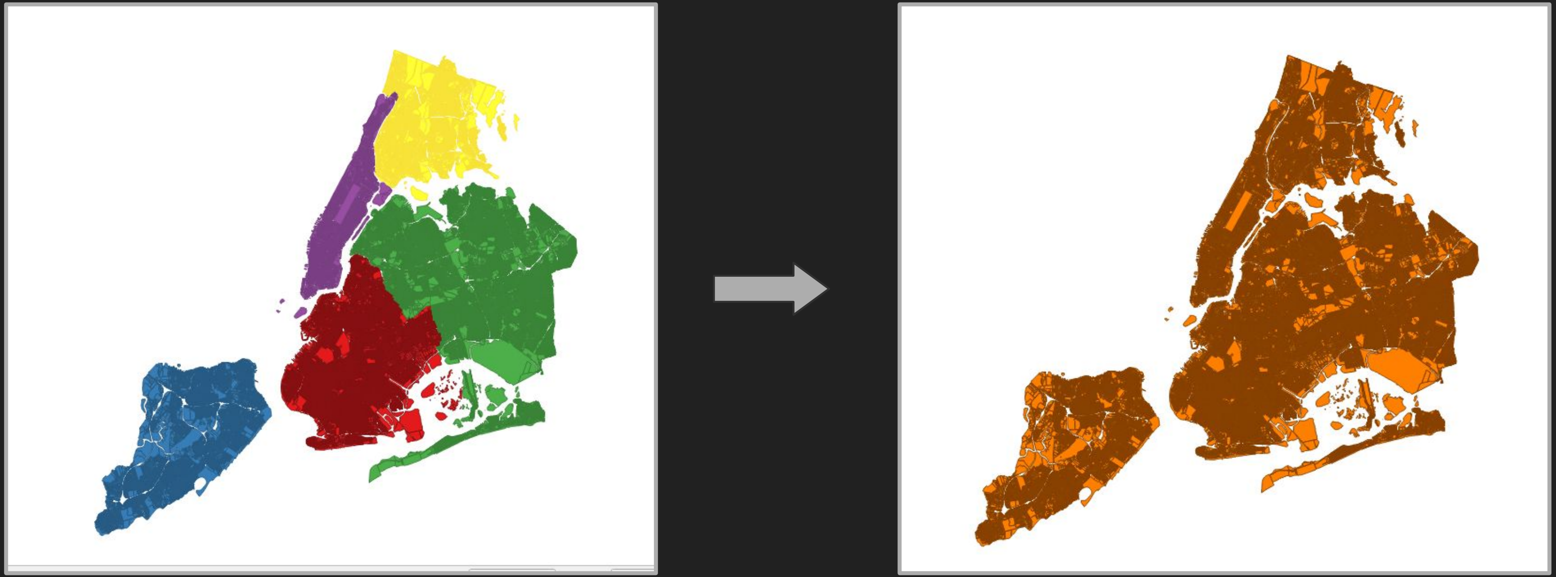
PART TWO

Progress in Our Pilot Study

Loading & Processing MapPLUTO (2002-2021)

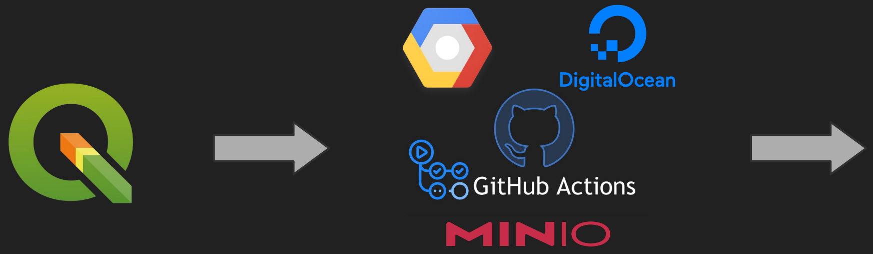
- Merging borough-year shapefiles

Merging Each Year’s Boroughs



*MapPLUTO_09v1 from BYTES

Loading & Processing MapPLUTO (2002-2021)

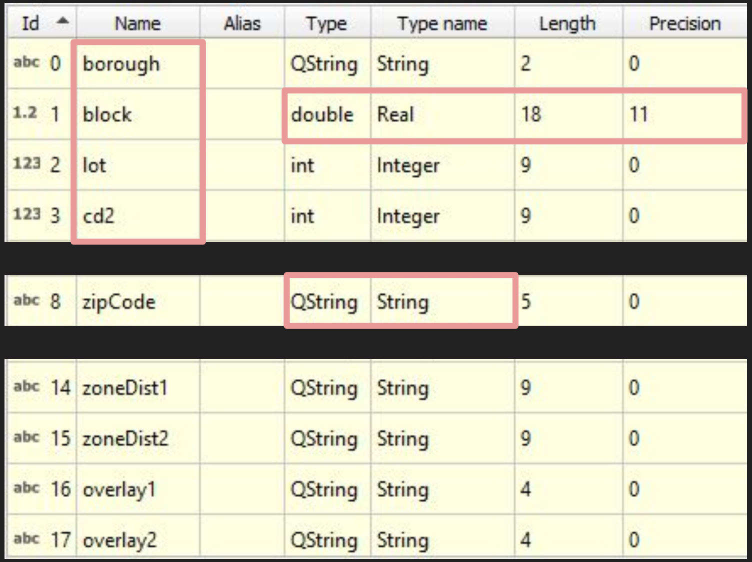


- Merging borough-year shapefiles

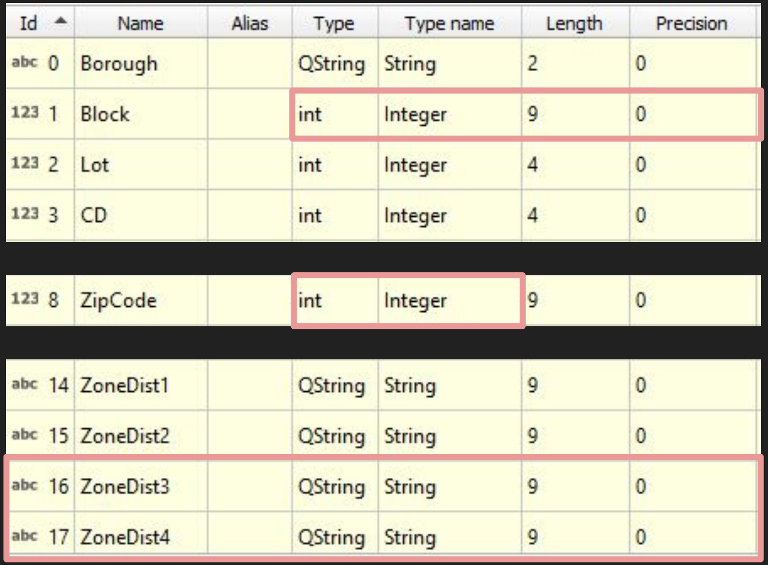
- Editing schemas
- Uploading data and metadata to BigQuery

Data Engineering pipeline

Schema Inconsistencies Between 2002 & 2009

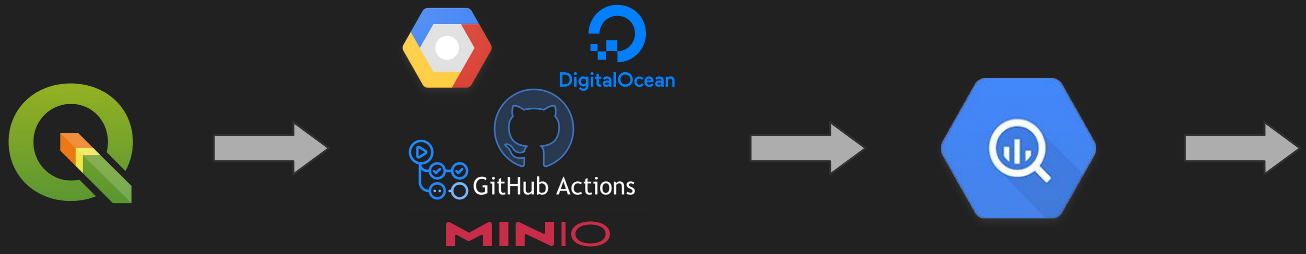


Id	Name	Alias	Type	Type name	Length	Precision
abc 0	borough		QString	String	2	0
1.2 1	block		double	Real	18	11
123 2	lot		int	Integer	9	0
123 3	cd2		int	Integer	9	0
abc 8	zipCode		QString	String	5	0
abc 14	zoneDist1		QString	String	9	0
abc 15	zoneDist2		QString	String	9	0
abc 16	overlay1		QString	String	4	0
abc 17	overlay2		QString	String	4	0



Id	Name	Alias	Type	Type name	Length	Precision
abc 0	Borough		QString	String	2	0
123 1	Block		int	Integer	9	0
123 2	Lot		int	Integer	4	0
123 3	CD		int	Integer	4	0
123 8	ZipCode		int	Integer	9	0
abc 14	ZoneDist1		QString	String	9	0
abc 15	ZoneDist2		QString	String	9	0
abc 16	ZoneDist3		QString	String	9	0
abc 17	ZoneDist4		QString	String	9	0

Loading & Processing MapPLUTO (2002-2021)



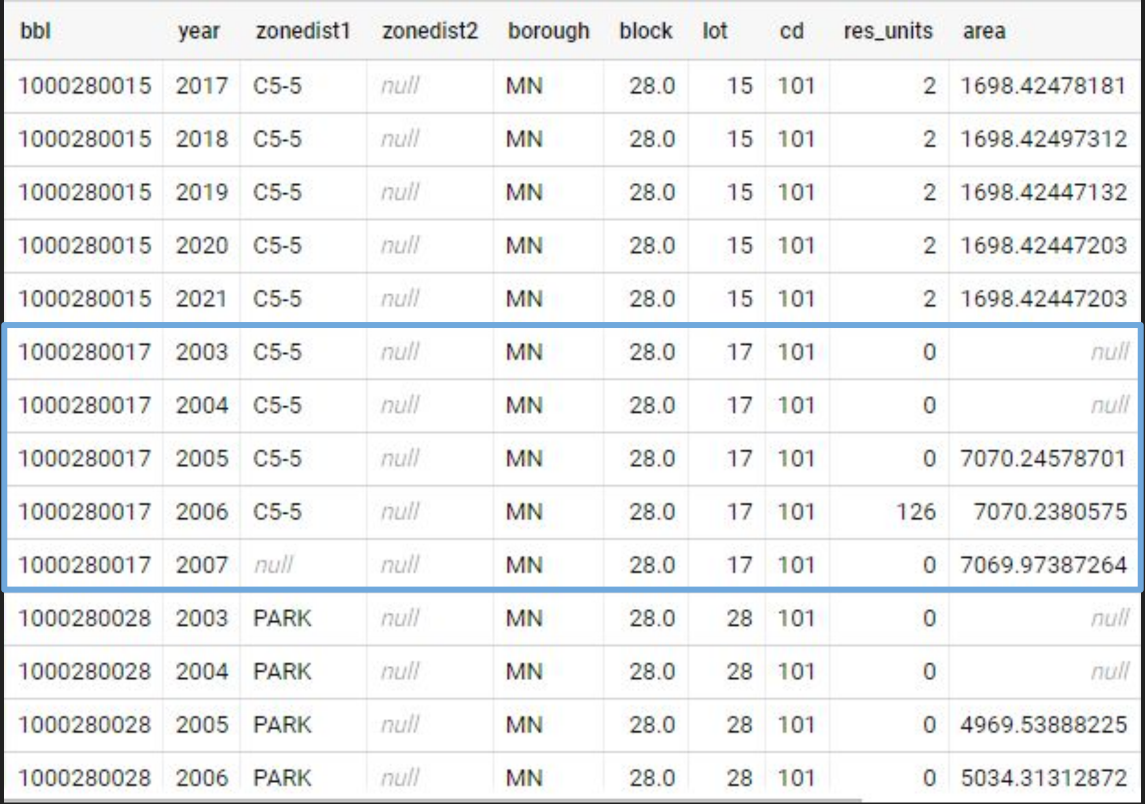
- Merging borough-year shapefiles

- Editing schemas
- Uploading data and metadata to BigQuery

Data Engineering pipeline

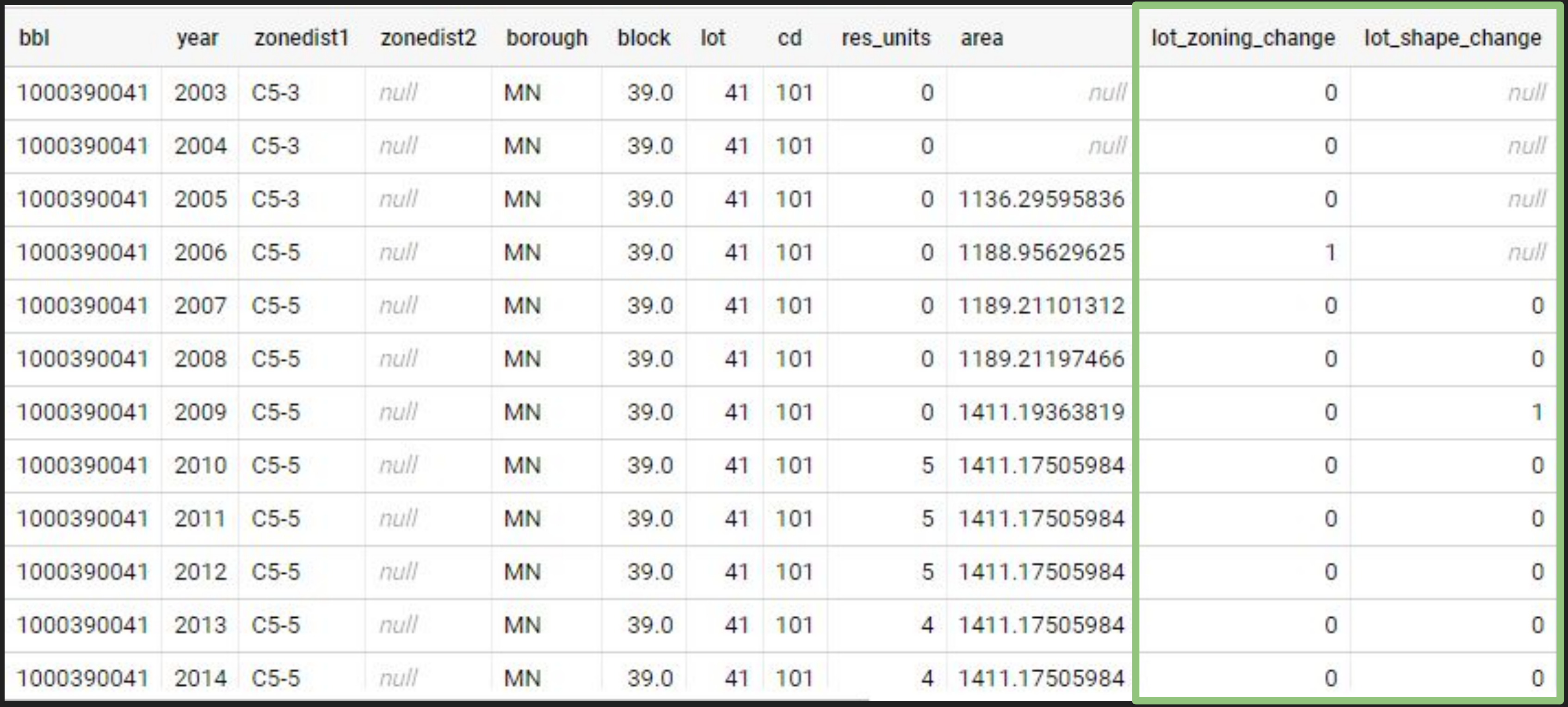
- Joining datasets ← **2GB in 15 seconds!!!**
- Creating new fields to track zoning changes
- Creating SQL views

Joining on BBL works well, but it isn’t perfect



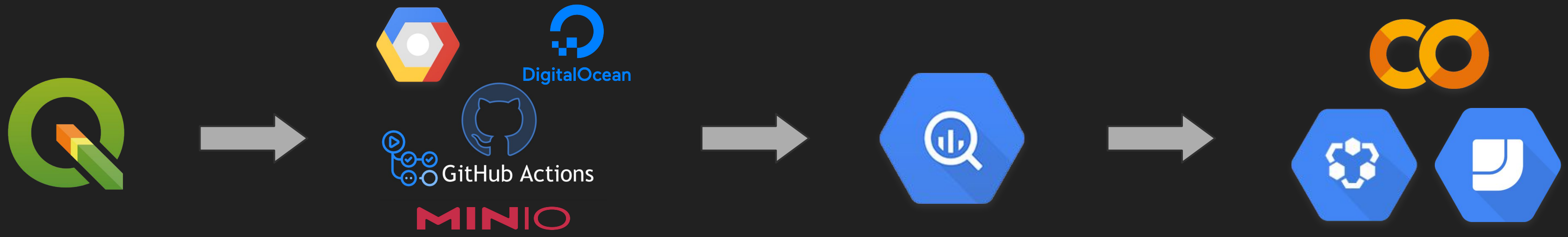
bbl	year	zonedist1	zonedist2	borough	block	lot	cd	res_units	area
1000280015	2017	C5-5	*null*	MN	28.0	15	101	2	1698.42478181
1000280015	2018	C5-5	*null*	MN	28.0	15	101	2	1698.42497312
1000280015	2019	C5-5	*null*	MN	28.0	15	101	2	1698.42447132
1000280015	2020	C5-5	*null*	MN	28.0	15	101	2	1698.42447203
1000280015	2021	C5-5	*null*	MN	28.0	15	101	2	1698.42447203
1000280017	2003	C5-5	*null*	MN	28.0	17	101	0	*null*
1000280017	2004	C5-5	*null*	MN	28.0	17	101	0	*null*
1000280017	2005	C5-5	*null*	MN	28.0	17	101	0	7070.24578701
1000280017	2006	C5-5	*null*	MN	28.0	17	101	126	7070.2380575
1000280017	2007	*null*	*null*	MN	28.0	17	101	0	7069.97387264
1000280028	2003	PARK	*null*	MN	28.0	28	101	0	*null*
1000280028	2004	PARK	*null*	MN	28.0	28	101	0	*null*
1000280028	2005	PARK	*null*	MN	28.0	28	101	0	4969.53888225
1000280028	2006	PARK	*null*	MN	28.0	28	101	0	5034.31312872

Creating New Fields to Track Zoning Changes



bbl	year	zonedist1	zonedist2	borough	block	lot	cd	res_units	area	lot_zoning_change	lot_shape_change
1000390041	2003	C5-3	null	MN	39.0	41	101	0	null	0	null
1000390041	2004	C5-3	null	MN	39.0	41	101	0	null	0	null
1000390041	2005	C5-3	null	MN	39.0	41	101	0	1136.29595836	0	null
1000390041	2006	C5-5	null	MN	39.0	41	101	0	1188.95629625	1	null
1000390041	2007	C5-5	null	MN	39.0	41	101	0	1189.21101312	0	0
1000390041	2008	C5-5	null	MN	39.0	41	101	0	1189.21197466	0	0
1000390041	2009	C5-5	null	MN	39.0	41	101	0	1411.19363819	0	1
1000390041	2010	C5-5	null	MN	39.0	41	101	5	1411.17505984	0	0
1000390041	2011	C5-5	null	MN	39.0	41	101	5	1411.17505984	0	0
1000390041	2012	C5-5	null	MN	39.0	41	101	5	1411.17505984	0	0
1000390041	2013	C5-5	null	MN	39.0	41	101	4	1411.17505984	0	0
1000390041	2014	C5-5	null	MN	39.0	41	101	4	1411.17505984	0	0

Loading & Processing MapPLUTO (2002-2021)



- Merging borough-year shapefiles

- Editing schemas
- Uploading data and metadata to BigQuery

Data Engineering pipeline

- Joining datasets
- Creating new fields to track zoning changes
- Creating SQL views

- Data exploration
- Analysis
- Visualization
- Collaboration

Working with 16 Millions Rows of MapPLUTO Data

We explored the data using Google Data Studio and Google CoLab

- In Google Cloud Platform
- Handles all the data without breaking a sweat
- Quick & Interactive visualizations

- Not part of Google Cloud Platform
- More complex exploratory analysis using Python
- Create smaller views to improve import times

Share GDS dashboard

PART THREE
Conclusion

Performance Evaluation

App	Data Size	Run Time	Ease of Use / Accessibility	Collaboration	Sharing Results	Recommended Skills	Drawbacks
BigQuery (storage)	⭐⭐⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	N/A	SQL	Many windows in a cramped UI.
Data Studio (interactive visuals)	⭐⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	N/A	Default settings can be annoying. Does not have all functions of other BI tools.
Colab (advanced analysis, visuals)	⭐⭐⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐	⭐⭐⭐	⭐⭐⭐⭐	SQL, Python	Version control and UX when collaborating needs improving.
Notebook (advanced analysis, visuals)	⭐⭐⭐	⭐⭐⭐	⭐⭐⭐	⭐	⭐⭐⭐⭐	SQL, Python	Slower run times and almost no collaboration.

Learn More

Read more about our overall experience and the Google Cloud Platform pilot

Thank you!

- What questions do you have about our work?
- What ideas do you have for using GCP in your work?