Planning in Crisis:

CTA Decision-Making Tools

Chicago | June 13, 2020

MScA Advisors

Agenda

Introduction

Technical Overview

Research Findings & Tools

Recommendations

MScA Advisors

Executive Summary

The **Chicago Transit Authority (CTA)** is a pillar of Chicago public life, completing over 450 million annual rides and employing more than 11,000 people. It has and must continue to be an essential service for Chicagoans.

The **CTA is facing unprecedented challenges** due to the economic fallout from COVID-19. Social distancing has necessitated a dramatically reduced ridership compared to previous years. This has left the CTA far below their expected revenue, a trend which is expected to continue in coming months.

At the same time, organizations are encountering new and reinforced expectations of social justice and fairness. This should be a consideration when changing services.

We have constructed a data pipeline and minimal viable product (MVP) which will allow CTA to understand and respond to this new environment.

We compiled essential data that will allow the CTA to appropriately allocate resources moving forward, while considering both policy objectives and the shareholder expectations.

The **MVP of the database and analytical tool** are built on zip-code level analyses of bus and train ridership, socio-economic indicators, COVID-19 cases, employment and public events. We have also provided suggestions for future iterations of this product.

MScA Advisors

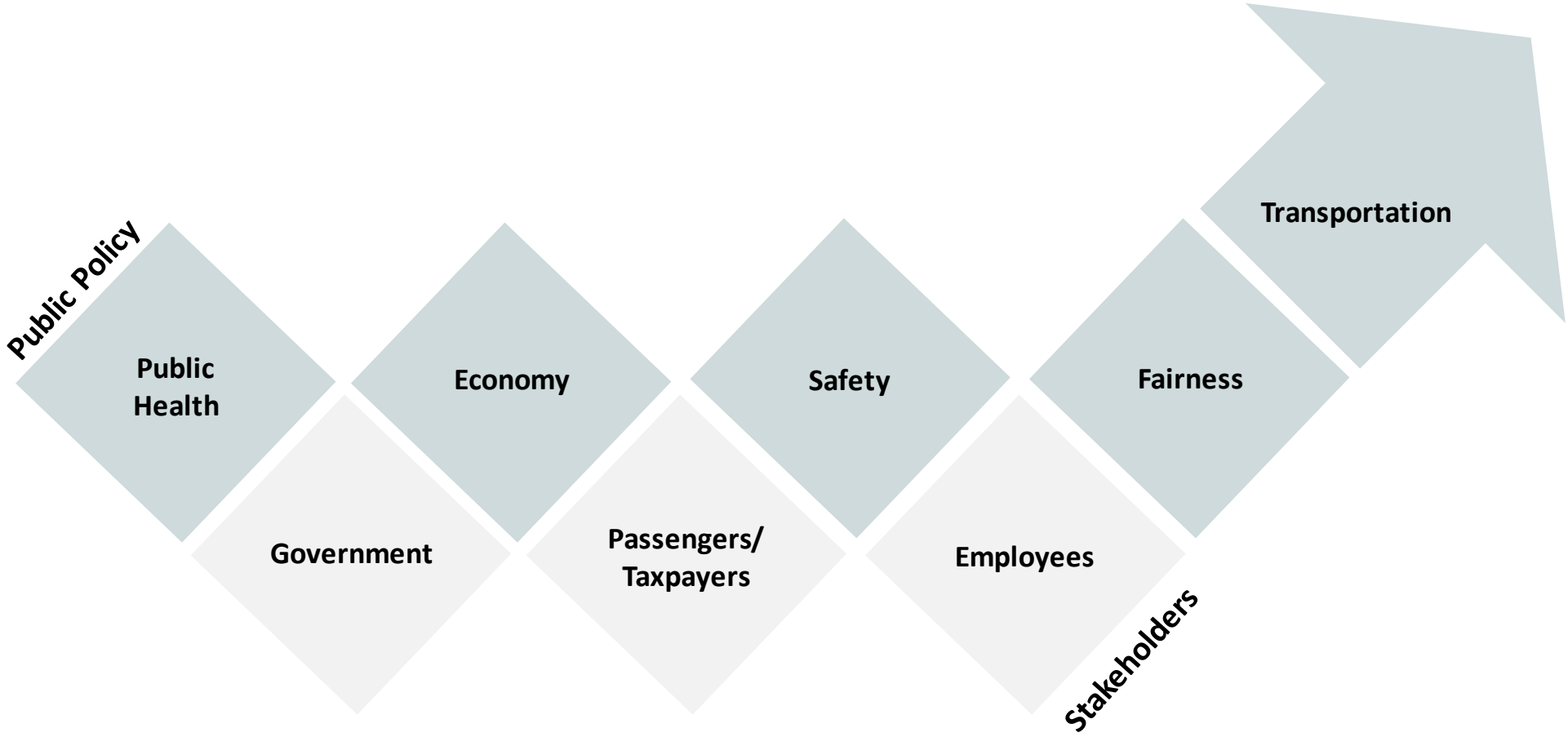
Business Case

Audience: CTA board & Chief Data Officer

Problem: A lack of **relevant aggregated** data leads to inefficient decision-making during crisis

Deliverable: MVP of a data pipeline and analytical tool, which will allow for prompt decision-making

Analytical Framework



MScA Advisors

Technical Overview

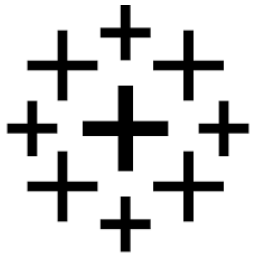
Recommendations

Introduction

Research Findings & Tools

MScA Advisors

Tools



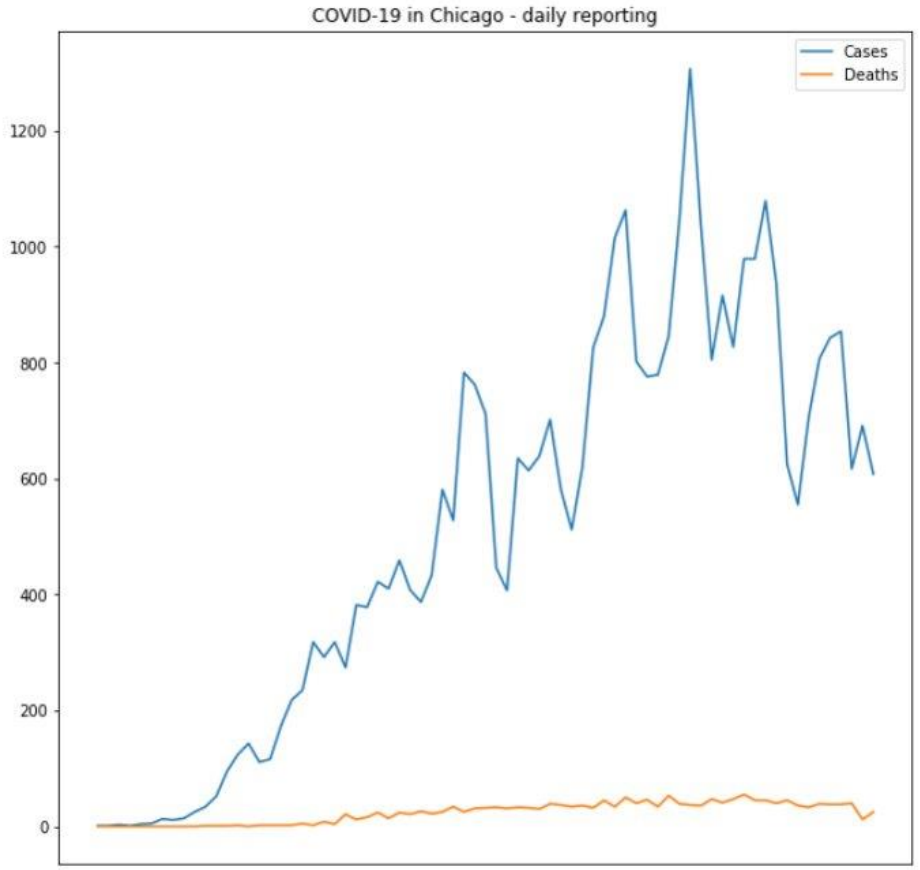
Data Profile

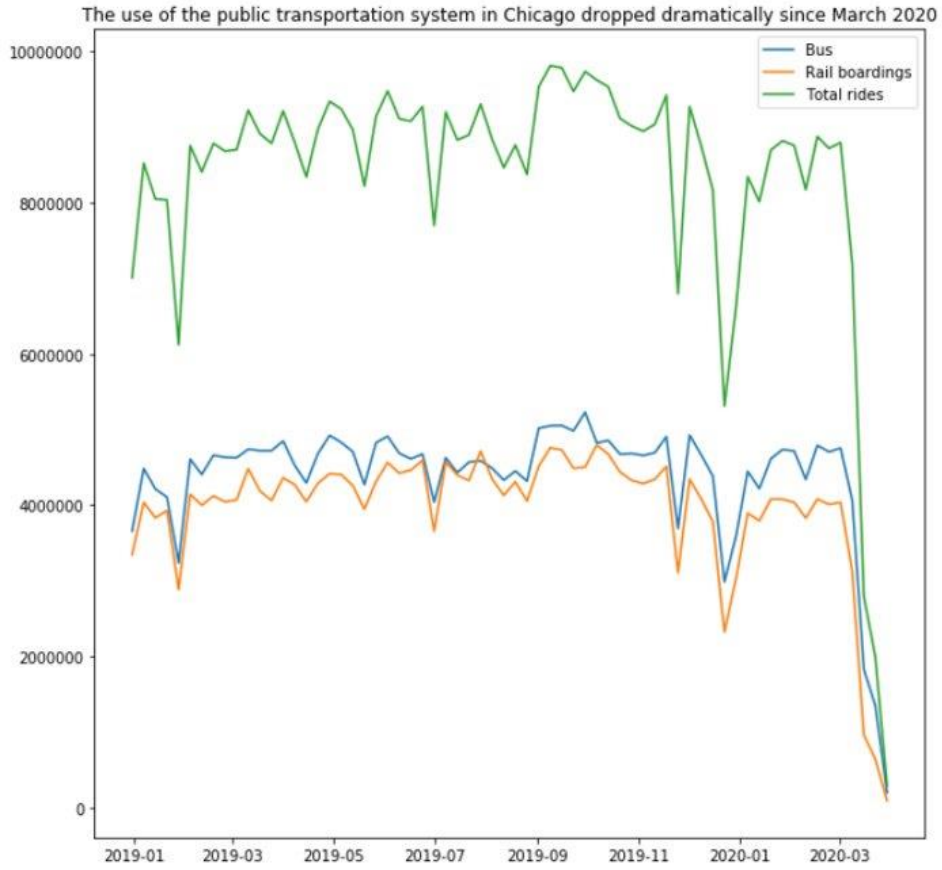
- **Outliers:** No
- **Anomalies:** Some Data Missing
- **Aggregations:** Multiple Joined Sources
- **Granularity:** Higher on main Fact Table than other data tables
- **Matching Methods/Algorithms:**
  - **Data cleaning:** Long/Lat & Zip codes
  - **Database management:** Primary/Foreign/Composite keys

Datasets: Data Quality Dimensions

	Data Size	Observations	Data Type	Missing Values	Validity	Uniqueness	Consistency	Timeliness	Accuracy
Google Mobility	20.1 MB	308237 * 10	Structured	683382	Yes	Yes	Yes	Yes(05/02/2020)	Yes
Chicago COVID-19 Statistics	12 KB	508 * 4	Structured	508	Yes	Yes	Yes	Yes(05/19/2020)	Yes
CTA Ridership Daily Boarding Totals	242 KB	7093 * 5	Structured	0	Yes	Yes	Yes	Yes(03/31/2020)	Yes
CTA Ridership Station Entries Daily Totals	35.1 MB	1001872 * 5	Structured	0	Yes	Yes	Yes	Yes(03/31/2020)	Yes
CTA Ridership Bus Routes Daily Totals	17.8 MB	855751 * 4	Structured	0	Yes	Yes	Yes	Yes(05/19/2020)	Yes
Employment	426 KB	42 * 238	Structured	*	Yes	Duplication**	Yes	Yes (03/2019)	Yes
Weather Google BigQuery - EPA	426 KB	1095	Structured	0	Yes	Duplication** **	Yes	Yes (3/2020)	Yes
Chicago events	9 MB	16741*9	Structured	***	Yes	Yes	Yes	Yes (05/21/2020)	Yes

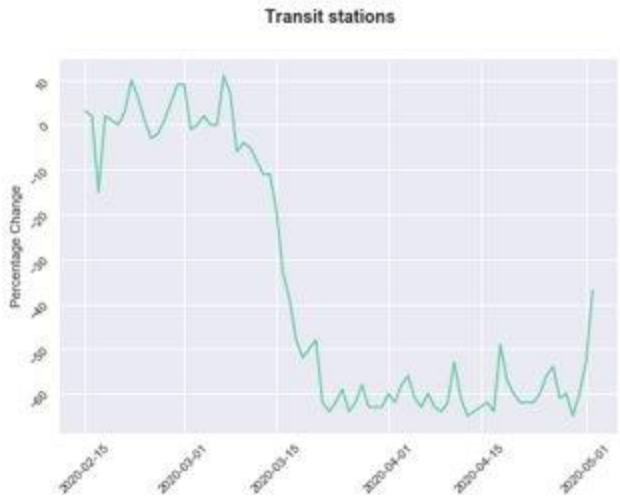
Drivers for Exploratory Analysis

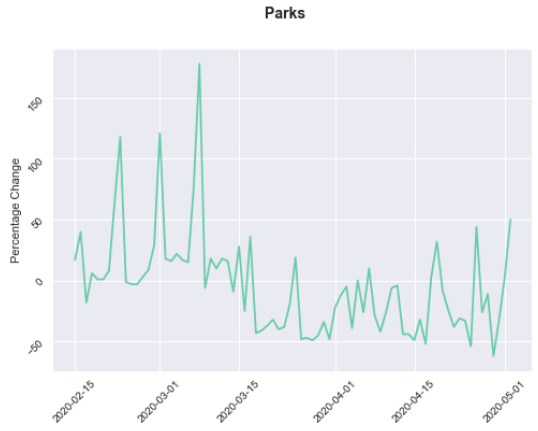


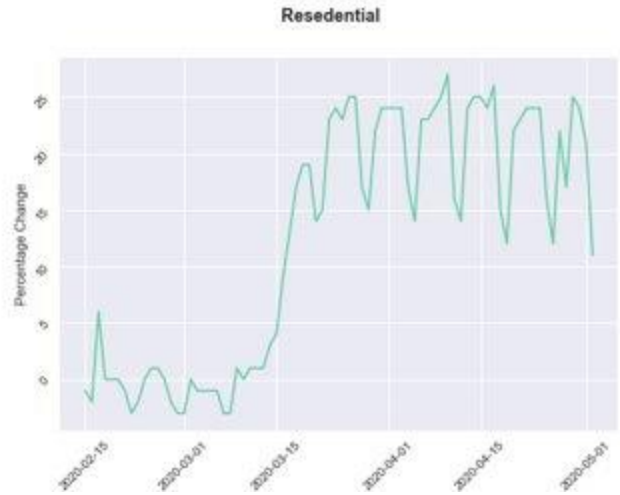


MScA Advisors

Exploratory Analysis







MScA Advisors

Data Cleaning

- **Tools:** Python (pandas) & R (dplyr)
- **Traditional challenges:** reshape/remove NAs/rename/join
- **Specific challenge:** assign zip codes, based on long/lat (from shapefile)

Clean and join

	idCTA_Sta	station_id	station_na	date	rides	longitude	latitude	zip_code
2	100001	40010	Austin-For	1/1/2019	576	-87.7768	41.87085	60304
3	100002	40010	Austin-For	1/2/2019	1457	-87.7768	41.87085	60304
4	100003	40010	Austin-For	1/3/2019	1543	-87.7768	41.87085	60304
5	100004	40010	Austin-For	1/4/2019	1621	-87.7768	41.87085	60304
6	100005	40010	Austin-For	1/5/2019	719	-87.7768	41.87085	60304

	STOP_ID	Location
2	30162	(41.857908, -87.669147)
3	30161	(41.857908, -87.669147)
4	30022	(41.829353, -87.680622)
5	30023	(41.829353, -87.680622)
6	30214	(41.831677, -87.625826)

	station_id	stationname	date	rides
2	40850	Library	10/9/2004	1057
3	40780	Central Park	6/18/2010	1154
4	41500	Montrose-Brown	10/30/2001	2116
5	40500	Washington/State	10/26/2006	0
6	41090	Monroe/State	7/7/2010	9431

Database Design Considerations

CAP Considerations	• Consistency • Partition Tolerance
Type of Application	• Analytical Focus: OLAP • Not used for any transactions
Data Format/Size	• Current: Structured, Publicly Avaliable • Future: Unstructured, Proprietary
Data Maintenance	• Initially Managed by MScA Advisors • Upgrade to automated end-to-end process
Hardware/Software	• Focus on ability to set-up quickly and scale • Future: Ability to handle Unstructured Data
Users	• 10-30 Users / Data Governance • Data Analyst to Senior Executive
Schedule/Budget	• Quickly Implement, before another crisis • Assume High CTA Priority and Budget Available

CURRENT

Distributed, Non-Partitioned Replicated, OLAP, Relational Database

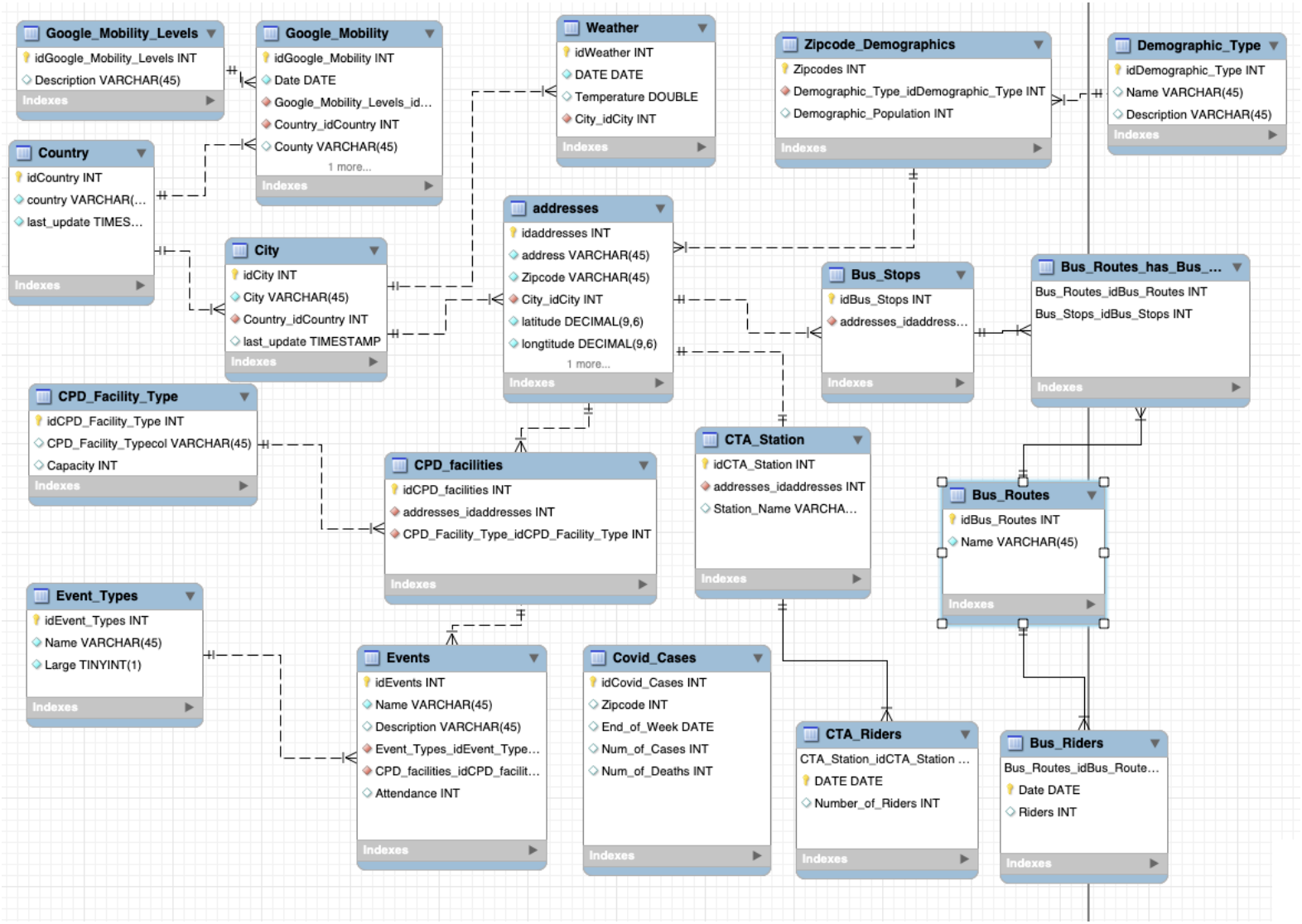
POTENTIAL FUTURE

Distributed, Partitioned Replicated, OLAP, Relational Database
&
Data Lake for Unstructured Data

Data Storage

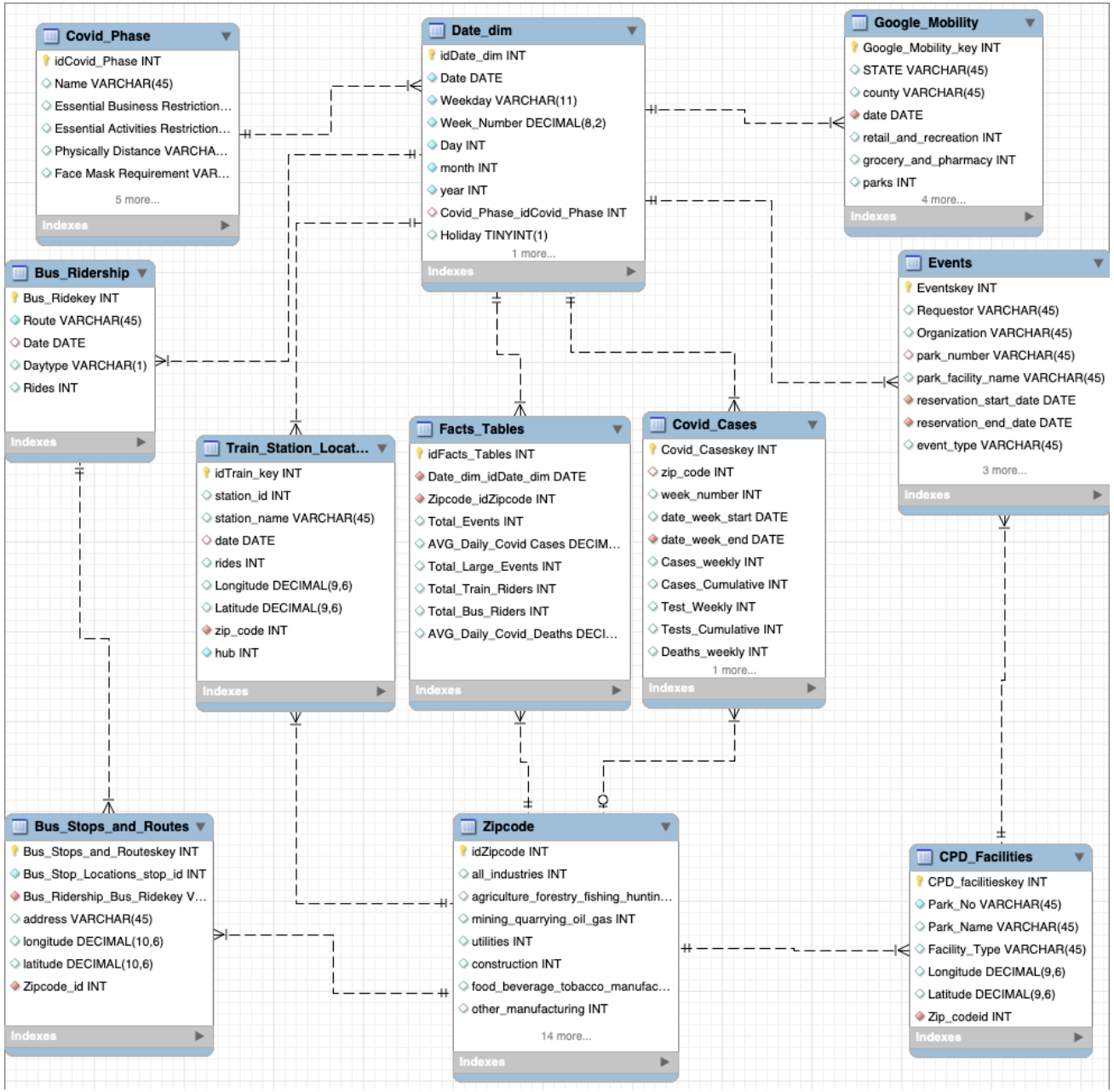
- **Key Considerations:**
  - Need a fast set-up
  - Ability to scale quickly in case we introduce new and unstructured data.
  - Low Maintenance Initial Set-up
- **Cloud Service – Choice: GCP**

Enhanced Entity-Relationship (OLTP)



MScA Advisors

Dimensional Model (OLAP)



MScA Advisors

NoSQL: Document-Oriented Database

Collection	Key	Type
Zipcodes		
	id	Int
	Demographic info	Object
	Demographic info	String
	geolocation	DECIMAL
	Covid Info	**Array**
	Date Id	Object
	Covid Cases	INT
	Covid Deaths	INT
	address	string
	Other Info	string

Collection	Key	Type
Train Stops		
	id	INT
	Name	string
	address	string
	Zipcode	INT
	longitude	Decimal
	latitude	Decimal
	Riders	**Array**
	id	Object
	Date	String
	# of Riders	INT

Collection	Key	Type
Parks		
	Park #	INT
	Park Name	string
	address	string
	ZipCode	INT
	longitude	Decimal
	latitude	Decimal
	Events	**Array**
	id	Object
	Date	String
	Attendance	INT
	Time	String
	Permit type	String

Collection	Key	Type
Bus Stops		
	id	INT
	Name	string
	address	string
	Zipcode	INT
	longitude	DECIMAL
	latitude	DECIMAL

Collection	Key	Type
Date		
	Day	String
	Weekday	String
	Holiday	String
	week	INT
	day	INT
	month	INT
	year	INT
	Temperature	Double
	News	**Array**
	id	Object
	Important News Story	String
	News Source	String
	Google Mobility Data	**Array**
	id County	Object
	Mobility Information	String

Collection	Key	Type
Bus Routes		
	Name	string
	Stops	Object
	Stop id	string
	Riders	**Array**
	id	Object
	Date	String
	# of Riders	INT

Technical Overview

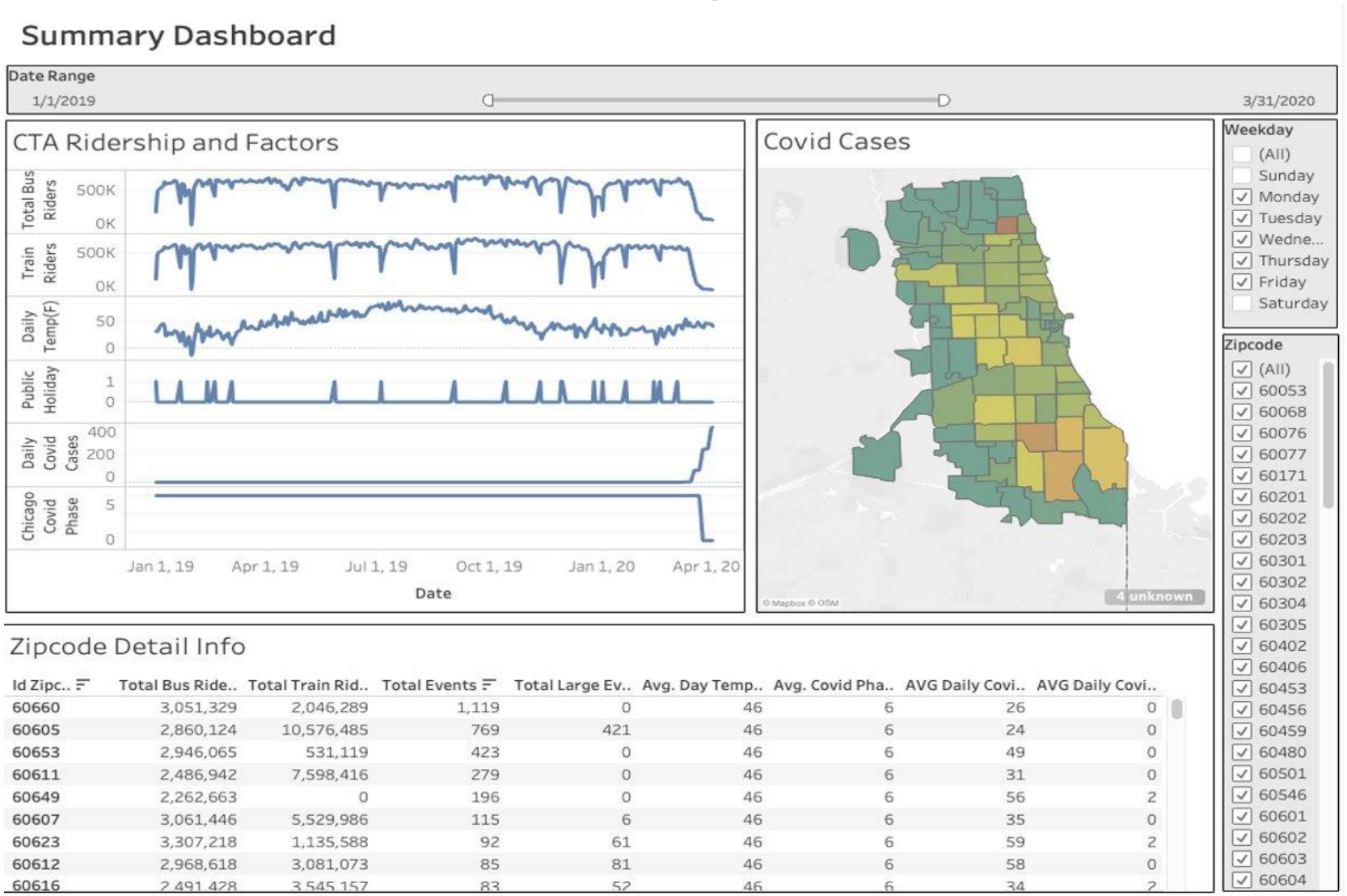
Recommendations

Introduction

Research Findings & Tools

MScA Advisors

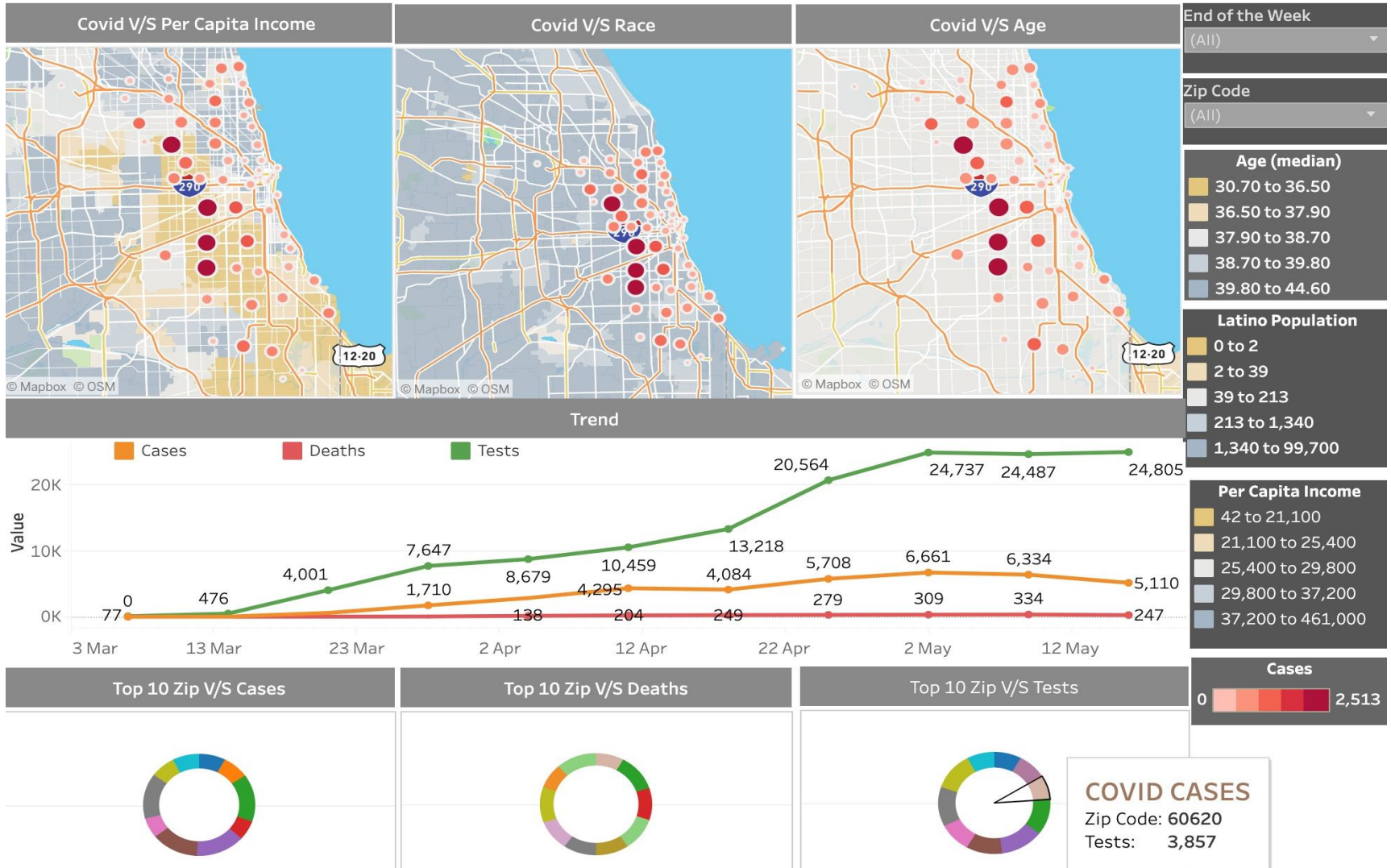
Insights and Summary Dashboard



Id Zipc..	Total Bus Ride..	Total Train Rid..	Total Events	Total Large Ev..	Avg. Day Temp..	Avg. Covid Pha..	AVG Daily Covi..	AVG Daily Covi..
60660	3,051,329	2,046,289	1,119	0	46	6	26	0
60605	2,860,124	10,576,485	769	421	46	6	24	0
60653	2,946,065	531,119	423	0	46	6	49	0
60611	2,486,942	7,598,416	279	0	46	6	31	0
60649	2,262,663	0	196	0	46	6	56	2
60607	3,061,446	5,529,986	115	6	46	6	35	0
60623	3,307,218	1,135,588	92	61	46	6	59	2
60612	2,968,618	3,081,073	85	81	46	6	58	0
60616	2,491,428	3,545,157	83	52	46	6	34	2

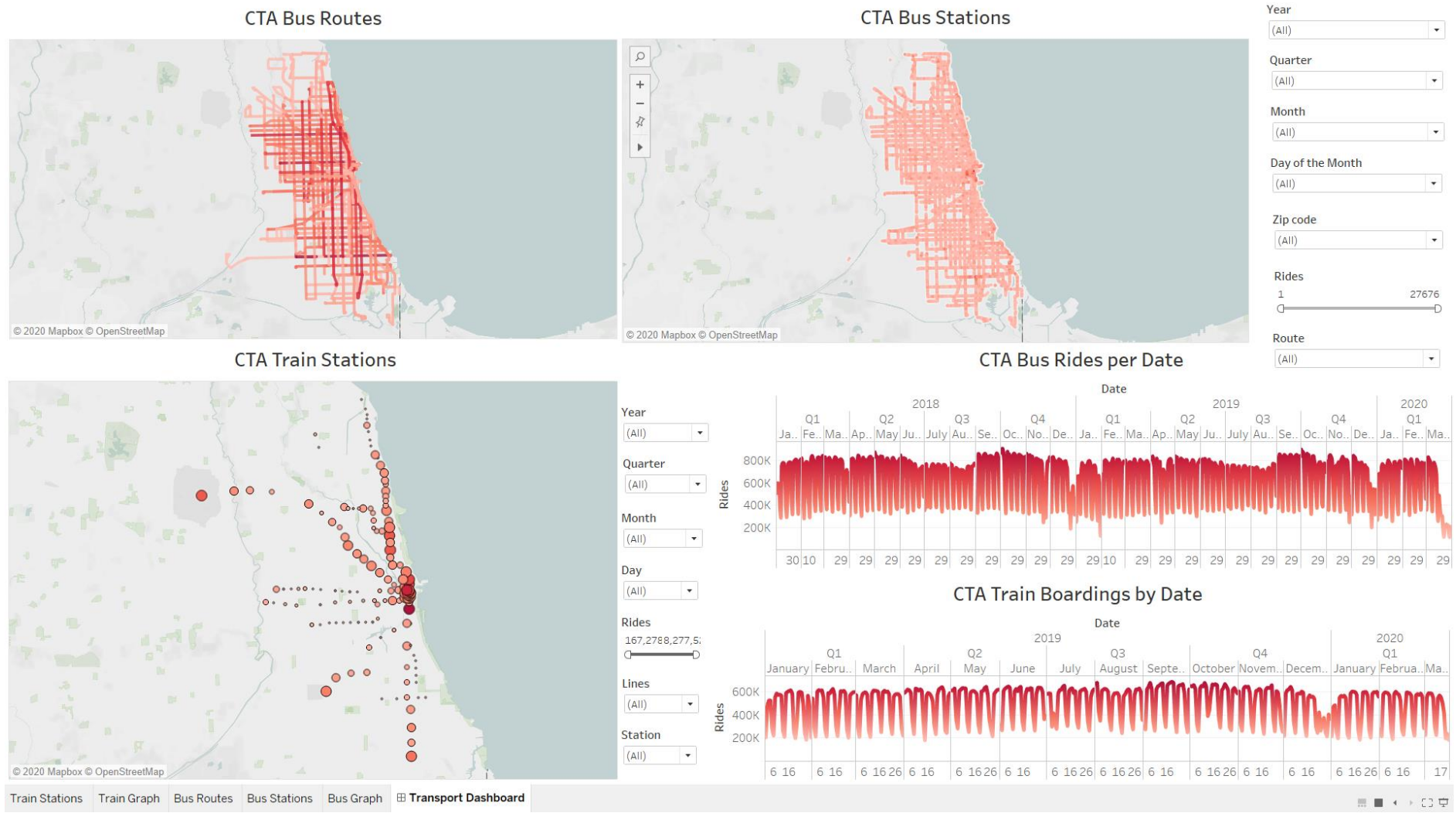
MScA Advisors

COVID-19 Analysis



MScA Advisors

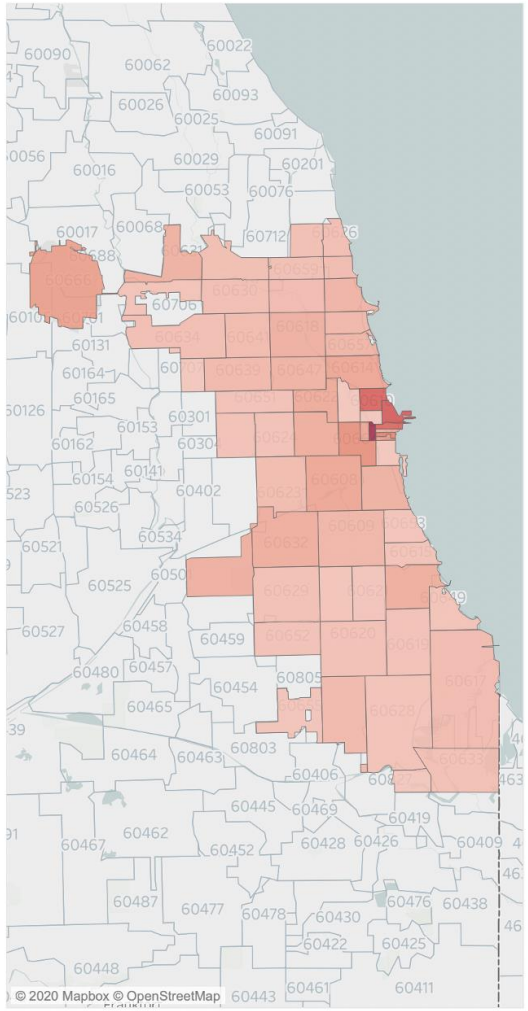
Transportation Analysis



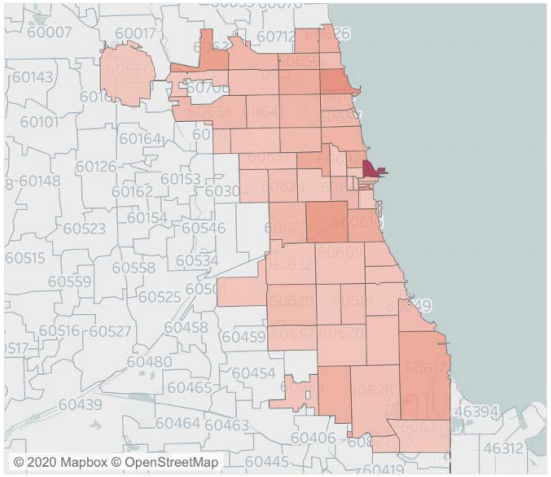
MScA Advisors

Employment Analysis

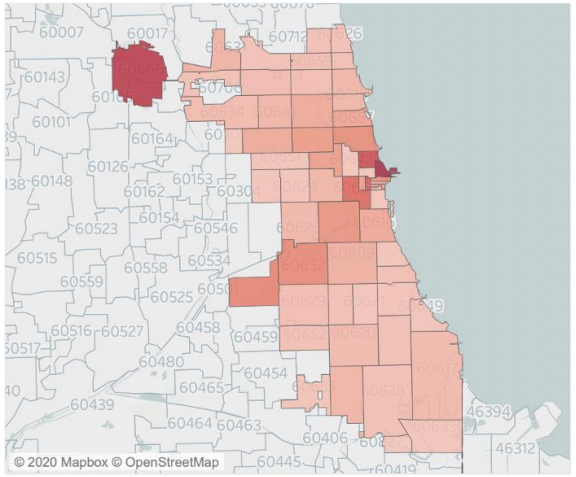
All Industries



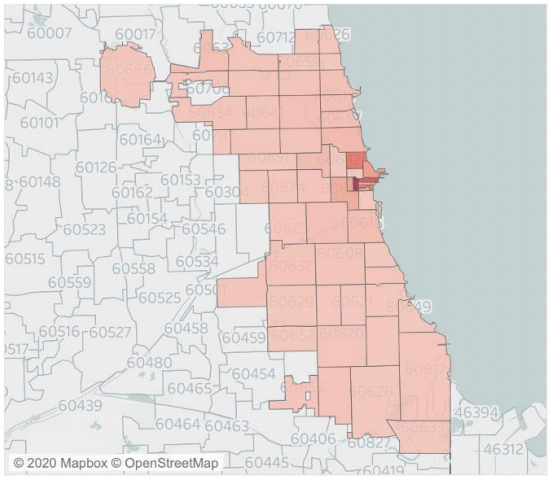
Healthcare



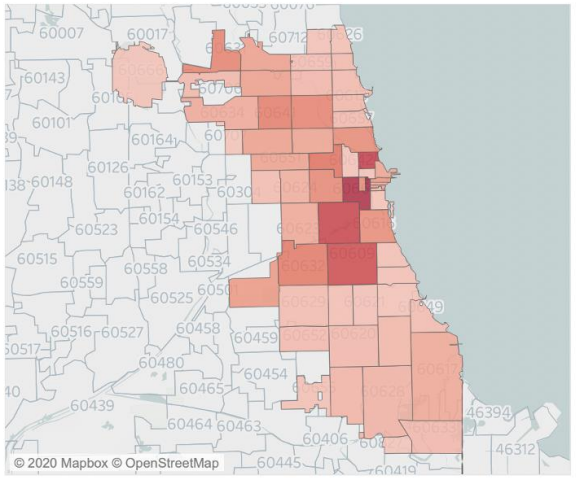
Essential Non-Health



Non-Essential Services



Non-Essential Trades



All Industries
0 139,075

Health & Social Services
0 21,806

Non-Health Essential Services
0 39,787

Non-Essential Trades
0 4,998

Non-Essential Services
0 88,912

Essential

Zip Code	
60611	61,593
60666	34,056
60610	32,663
60607	25,330
60606	24,783
60608	20,059
60614	17,916
60638	16,759
60601	16,574
60632	16,223
60622	15,662
60647	13,739

Non-Essential

Zip Code	
60606	93,910
60601	63,752
60610	49,097
60603	40,590
60611	26,518
60607	21,204
60661	21,036
60604	18,039
60602	17,189
60614	10,162
60622	9,707
60608	8,945

MScA Advisors

Predictive Analytics: Ideas and Challenges

Prophet is an open source science tool developed by the Data science team at Facebook for time series forecasting

Uses **Generalized additive models** which can be represented by

$$y(t) = g(t) + s(t) + h(t) + \epsilon$$

Where

$g(t)$= models trend

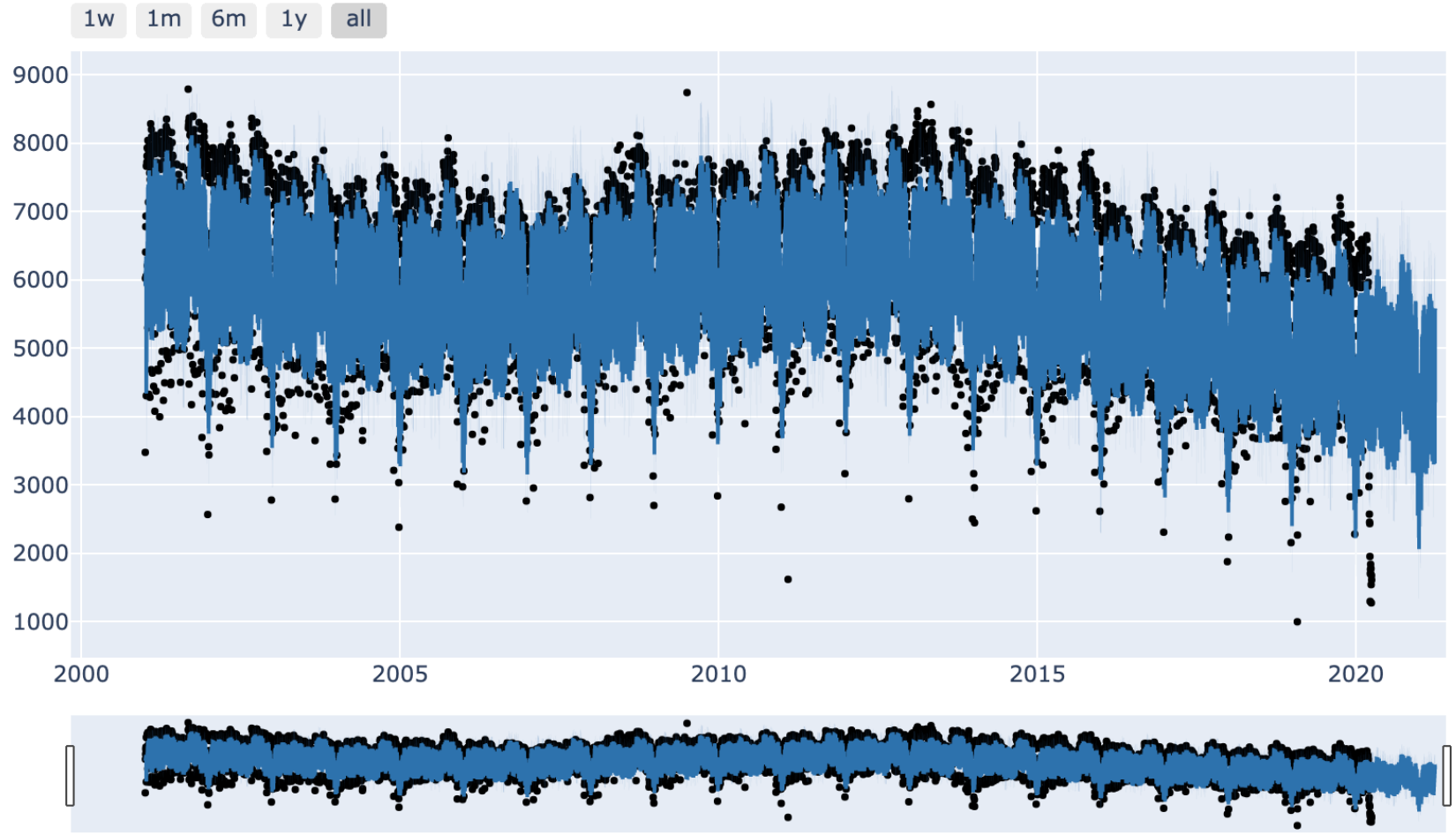
$s(t)$ = models seasonality

h(t)= models the effects of holidays

ϵ = error term

MScA Advisors

Bus Ridership Prediction



MScA Advisors

Introduction

Technical Overview

Research Findings & Tools

Recommendations

MScA Advisors

Policy Recommendations

Expect continued low ridership downtown as non-essential service workers continue to work from home

Consider reducing frequency of services downtown, on all lines/routes but the Orange/Blue lines

Continue health and safety measures, particularly in COVID hotspots

Explore opportunities for changed revenue structure

MScA Advisors

Technical Recommendations

Update transport data frequently for faster decisions

Share, engage, and collaborate with other organizations

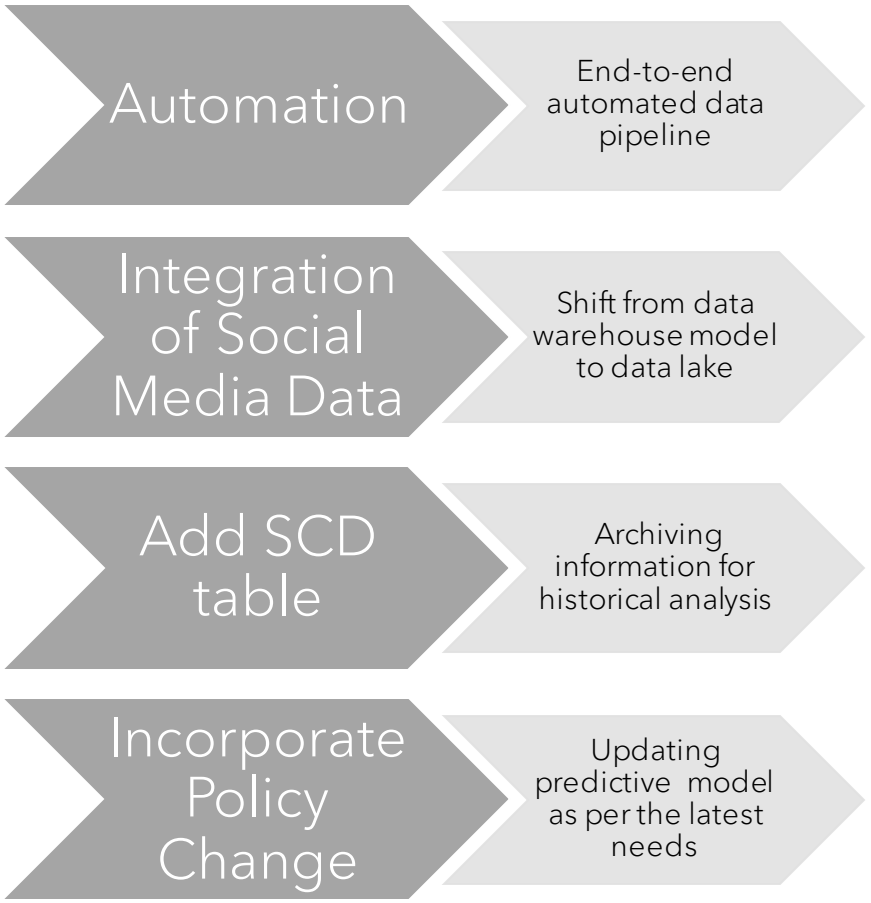
Prepare prediction models for crisis situations

Add other 3rd party providers to the product through APIs for adding a holistic big picture

Challenges

- Structure of data
- Compatibility of Spatial Files and Google Cloud SQL
- Frequency of Data
- Decentralized Version Control for Data

Future Work


Automation
End-to-end automated data pipeline
Integration of Social Media Data
Shift from data warehouse model to data lake
Add SCD table
Archiving information for historical analysis
Incorporate Policy Change
Updating predictive model as per the latest needs

Project Lessons

- **Cloud solutions** are better for interoperability and teamwork
- **One** programming language **is good for consistency**, but **multiple** ones can be **great complements for resolving bottlenecks**
- Proper database management **will take 50% of project time**
- **Communication** is as important as all **technical components**
- **Less tools** are better for **reproducibility** of work by others

Team Lessons

- **Proactive team formation** leads to satisfaction and success
- **Diversity in minds and backgrounds** is crucial for strong results
- **Commitment and accountability** is as important as tech skills
- **Proper planning and agile teamwork** is essential for timely steps
- **Micro-teams (2 people)** are better for resolving tech issues quickly

MScA Advisors

Chris Reimann
Co-Founder

Kyla Ronellenfitsch
Co-Founder

Oleksiy Anokhin
Co-Founder

Devanshi Verma
Co-Founder

THE UNIVERSITY OF CHICAGO

MScA Advisors

References

- GitHub Repository
  - https://bit.ly/depa-team-awesome
- Tableau Dashboards
  - https://bit.ly/transport-analysis
  - https://bit.ly/covid19-chicago-analysis
  - https://bit.ly/employment-analysis
  - https://bit.ly/cta-summary-analysis

Questions

MScA Advisors