8-2017

Integrating apache spark and R for big data analytics on solving geographic problems

Mengqi ZHANG
Singapore Management University, mqzhang.2015@mais.smu.edu.sg

Tin Seong KAM
Singapore Management University, tskam@smu.edu.sg

Integrating Apache Spark and R for Big Data Analytics on solving geographic problems

SINGAPORE MANAGEMENT UNIVERSITY

ZHANG MENGQI & PROFESSOR KAM TIN SEONG

Content

- Motivation & Objective
- Proposed Approach
- Data Descriptions
- Methodology
- Application Framework
- Demo
- Analysis & Results
- Conclusion & Future work

Motivation

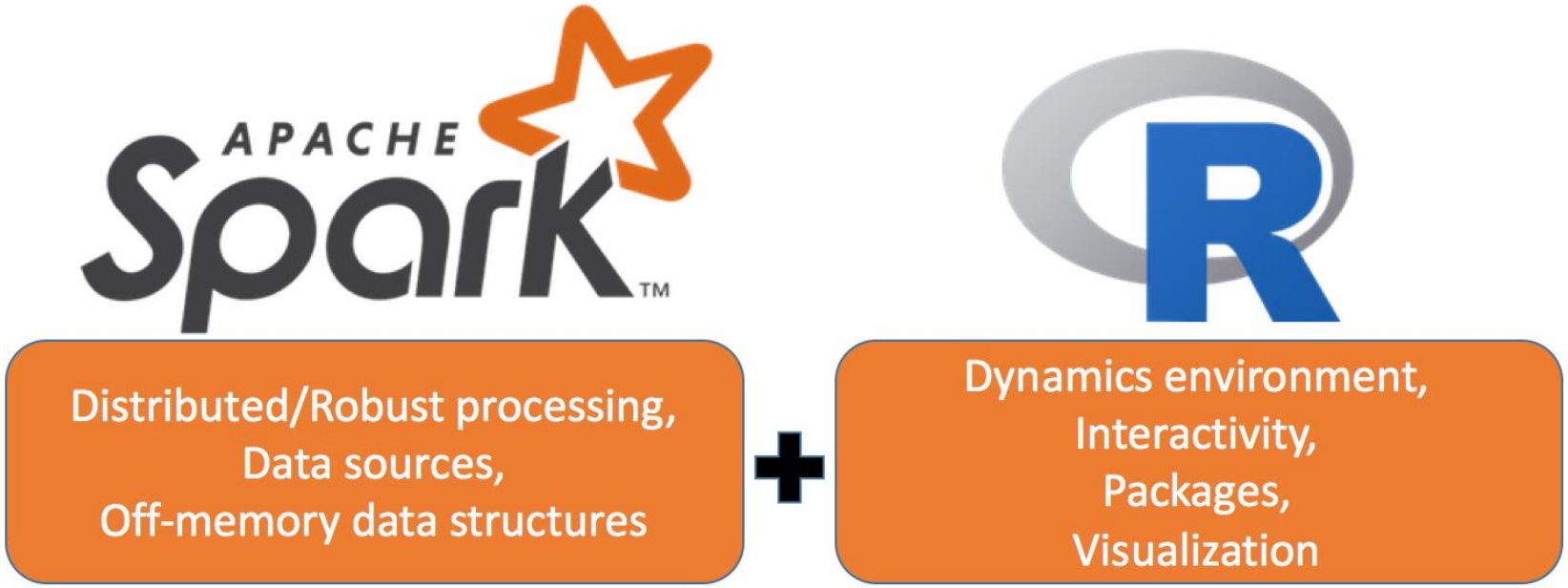
- A flood of digital data is being **generated every day**.
- The true values of these data will **not be fully appreciated** until they have been well processed and interpreted.
- The majority of the current academic research and practice development efforts tend **to focus on the technological aspect of big data instead of analytical aspect**.

Objectives

- To **design and implement a big data analytics application** by **integrating an open source big data processing framework** and **an open source data analysis environment**.
- To report our findings gain through working on this project
- To recommend future research difficulties based on our experiment

Proposed Approach

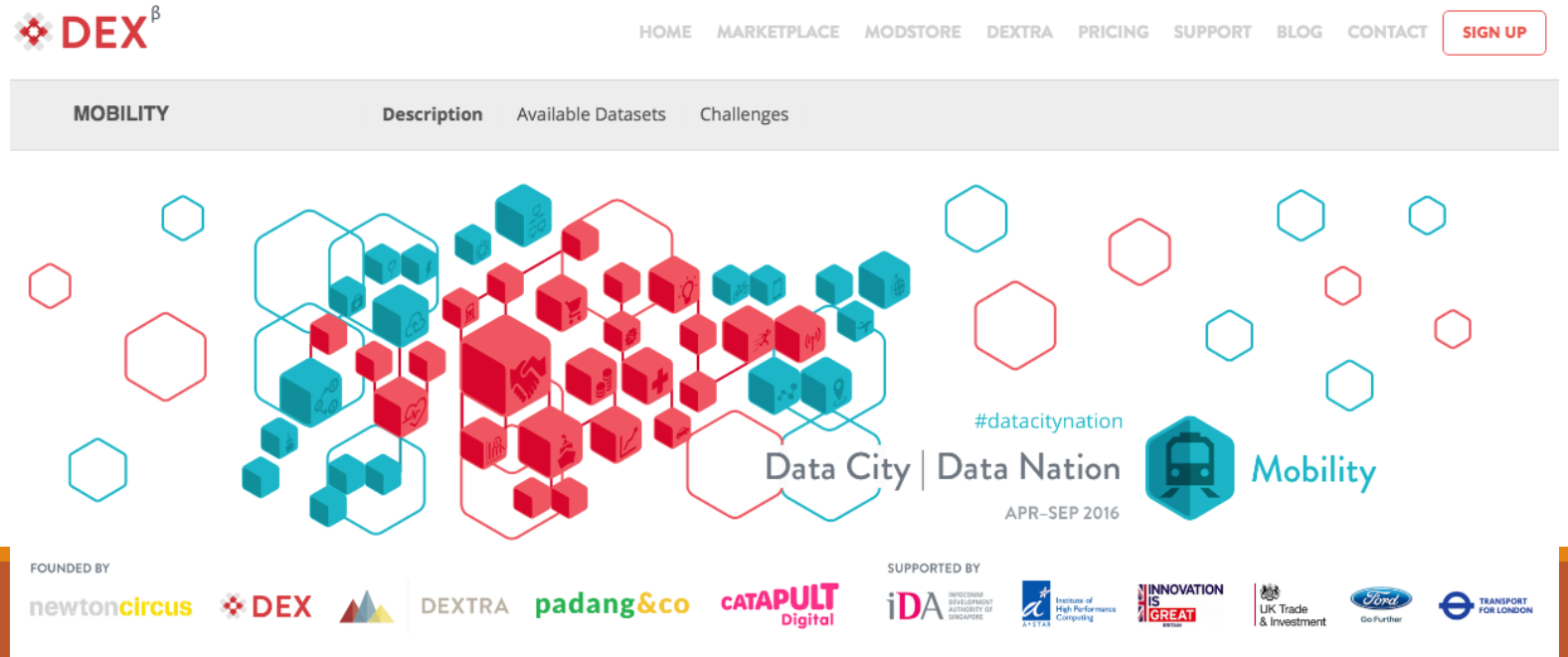
<u>Perform Big Data Analytics:</u> Integrate the existing big data technology with conventional data analysis tools and test it with real dataset.



Solution in search of Problem

Data City | Data Nation Mobility: “The possibilities are yours to discover. ”

- Uncover trends in how we go about our day, being on the move to live, work and play.
- Investigate our lifestyle choices and everyday actions of transportation.



Solution in search of Problem

- Study the spatial-temporal patterns of bus commuters within Singpaore's planning subzones and between planning subzones.
- Construct a spatial interaction model using commuters' data.

Data	Descriptions	Owner
Singapore Master Plan 2014 Subzone Boundary (No Sea) (SHP)	Provides indicative polygon of subzone boundary till 2014.	Urban Redevelopment Authority (URA) of Singapore
CITY_NATION_RIDE_DATA_FULL.csv	Provides the fare card data records across Singapore for 15/02/2016 00:01 to 21/02/2016 23:59.	Land Transport Authority (LTA) of Singapore
lta-bus_stop_locations_2011_09_05.csv	Provides the X,Y coordinates for Singapore's bus stops till 05/09/2011.	Land Transport Authority (LTA) of Singapore
lta-sbst_route_2011-09-05.csv	Provides the Singapore's bus service routes with bus stop codes till 05/09/2011.	Land Transport Authority (LTA) of Singapore

Data Descriptions: Datasets

- **Singapore Master Plan 2014 Subzone Boundary (No Sea) (SHP)**
  - 323 Subzones

Subzone matrix: 323 × 323.
Too Big to be analyzed
without Big Data Tech.

Figure. Polygon of Singapore's subzone boundary

Data Descriptions: Datasets

- **CITY_NATION_RIDE_DATA_FULL.csv**
  - Land Transport Authority (LTA) of Singapore
  - 50,697,538 records across Singapore (4.82GB)
  - for 15/02/2016 00:01 to 21/02/2016 23:59

CARD_ID	ACTUAL_SRVC_NUMBER	BOARDING_STOP_STN	ALIGHTING_STOP_STN	RIDE_START_DATE	RIDE_START_TIME	RIDE_END_DATE	RIDE_END_TIME
2570CDFC527F803775E32B1909A8210978604AF9	912	46088	46621	21/2/16	23:50:04	21/2/16	23:57:45
2E13336EFA135C4827771DEFCE8A5F117C88F721	912P	46088	46501	18/2/16	08:24:54	18/2/16	08:27:28
EF79FB78E2FF2D9912CEE93898156C8E3033933B	913	46088	46501	19/2/16	07:13:27	19/2/16	07:15:53

Data Descriptions: Datasets

- **lta-bus_stop_locations_2011_09_05.csv(4659 bus stops)**

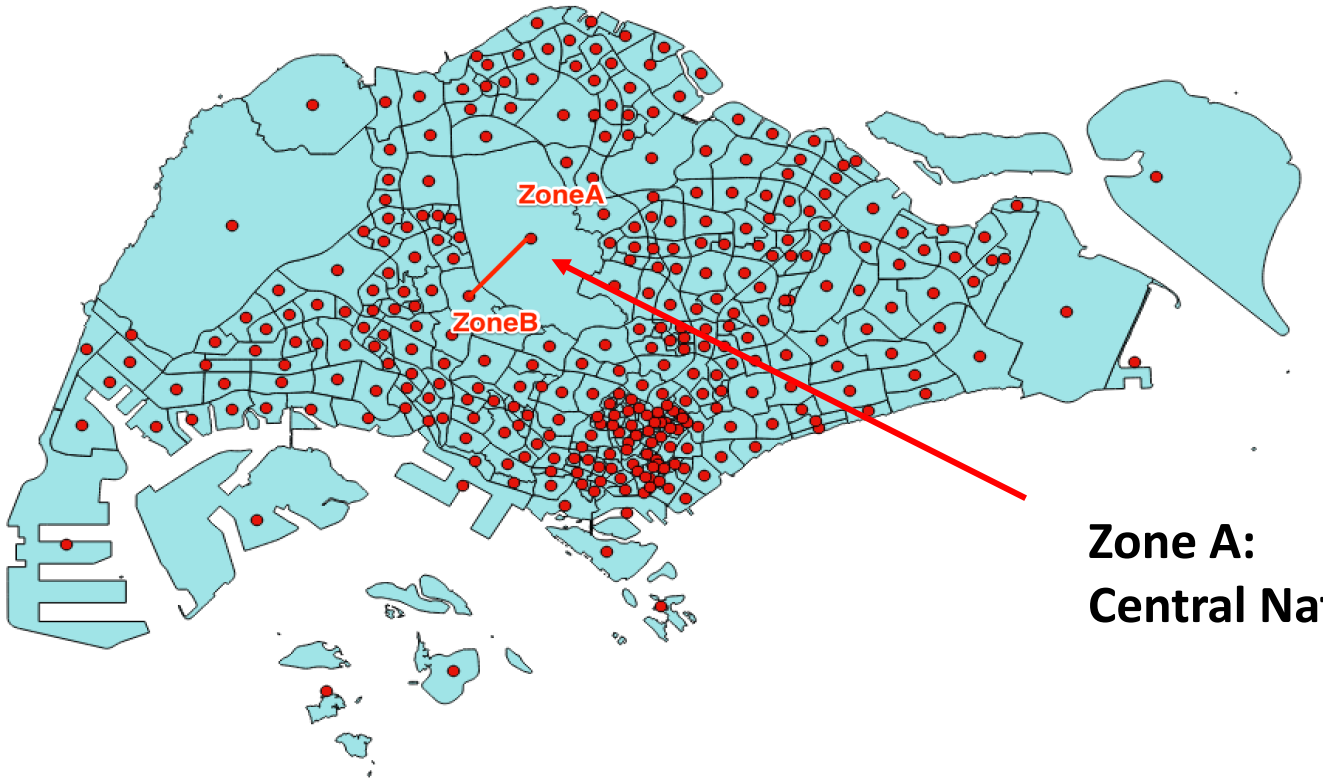
X	Y	ZID	NAME	ELEV	ICON
103.8525359	1.296848255	0	1012	0	81
103.8532247	1.297709706	1	1013	0	81
103.8530216	1.296982901	2	1019	0	81
103.8544142	1.296672985	3	1029	0	81

- **lta-sbst_route_2011-09-05.csv**

SR_SVC_NUM	SR_SVC_DIR	SR_ROUT_SEQ	SR_BS_CODE	SR_DISTANCE
106	1	1	43009	0
106	1	2	43179	0.6
106	1	3	43189	1
106	2	1	11401	0
106	2	2	11239	0.2
106	2	3	11229	0.5

Data Descriptions: Datasets

- **Distance calculation between Subzone pairs**
  - **Type 1: Euclidean Distance**
  - **Type 2: Bus route distance**



Zone A:
Central Natural Reserve.

Figure. Polygon of Singapore's subzone boundary with centroids

Data Descriptions: Datasets

- **Distance calculation: Distance Matrix(Bus route distance)**

		Destination							
		SERANGOON CENTRAL	KALLANG BAHRU	YISHUN EAST	YIO CHU KANG EAST	YIO CHU KANG	BAYSHORE	DEFU INDUSTRIAL PARK	
Origin	SERANGOON CENTRAL	0	7.57	999	13.68	14.92	13.59	4.02	
	KALLANG BAHRU	7.57	1e-6	999	999	11.33	12.86	999	
	YISHUN EAST	999	999	1e-6	8.49	999	999	999	
	YIO CHU KANG EAST	13.68	999	999	1e-6	6.28	18.5	8.38	
	YIO CHU KANG	14.92	11.33	NA	6.28	1e-6	21.52	9.70	
	BAYSHORE	13.59	12.86	999	18.5	21.52	1e-6	10.14	
	DEFU INDUSTRIAL PARK	4.02	999	999	8.38	9.70	10.14	1e-6	

Table. Bus route distance (km) between some subzones in Singapore

Data Descriptions: Datasets

- **Fare card flow aggregation**

		Destination							
		SERANGOON CENTRAL	KALLANG BAHRU	YISHUN EAST	YIO CHU KANG EAST	YIO CHU KANG	BAYSHORE	DEFU INDUSTRIAL PARK	Oi
Origin	SERANGOON CENTRAL	32588	995	832	315	559	25	1377	36691
	KALLANG BAHRU	590	1468	0	4	0	12	0	2074
	YISHUN EAST	468	0	19758	0	0	0	0	20226
	YIO CHU KANG EAST	568	4	0	1032	1995	2	37	3638
	YIO CHU KANG	919	140	194	24104	20159	17	74	45607
	BAYSHORE	152	3	0	0	16	1078	0	1249
	DEFU INDUSTRIAL PARK	673	0	0	0	142	29	4619	5463
	Dj	35958	2610	20784	25455	22871	1163	6107	

Table. Example for Flow matrix for the Origin and the Destination pairs

Methodology

- **Spatial Interaction Models:**
  - Predict the amount or likelihood of people or goods (or service or information) moving between two locations in space.
  - Example: Internal migration flows in Austria

		Destination									
		AT11	AT12	AT13	AT21	AT22	AT31	AT32	AT33	AT34	O_i
Origin	AT11	0	1131	1887	69	738	98	31	43	19	4016
	AT12	1633	0	14055	416	1276	1850	388	303	159	20080
	AT13	2301	20164	0	1080	1831	1943	742	674	407	29142
	AT21	85	379	1597	0	1608	328	317	469	114	4897
	AT22	762	1110	2973	1252	0	1081	622	425	262	8487
	AT31	196	2027	3498	346	1332	0	2144	821	274	10638
	AT32	49	378	1349	310	851	2117	0	630	106	5790
	AT33	87	424	978	490	670	577	546	0	569	4341
	AT34	33	128	643	154	328	199	112	587	0	2184
	D_i	5146	25741	26980	4117	8634	8193	4902	3952	1910	89575

Source: Eurostat – Table migr_r_2at

Methodology

- Apply spatial interaction model to discover the passengers' behaviors in a bus transportation system based on fare card data for Singapore.
- The doubly – constrained Wilson-style spatial interaction model for this transportation system takes the form:

$$P_{ij} = A_i B_j O_i D_j f(d_{ij}) \qquad (1)$$

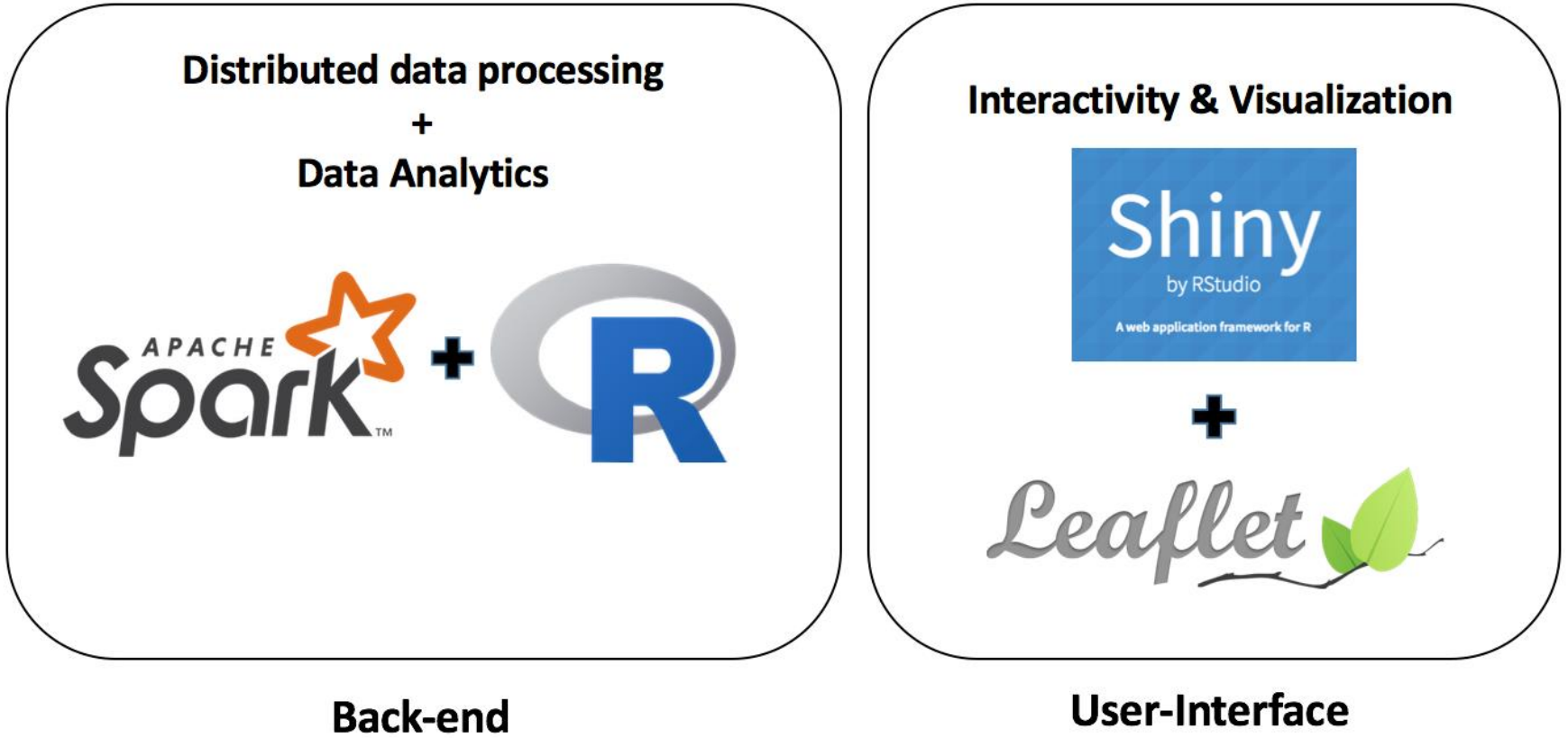
  - O_i is the total number of passengers that have left origin zone
  - D_j is the total number of passengers that have arrived at destination zone
  - $A_i$and B_j are balancing factors which ensure that the estimates of P_{ij}, when summed across both rows and columns of the matrix, equal the known O_i and D_j totals
  - $f(d_{ij})$term drives the estimates in the model and it represents the distance between O_i and D_j. In a transportation model

Application Framework

- **SparkR**
  - Provides a light-weight frontend to use Apache Spark from R.
  - Provides a distributed DataFrame implementation that supports operations like selection, filtering, aggregation etc. (similar to R data frames, dplyr) but on large datasets.
  - **Supports distributed machine learning using MLlib.**

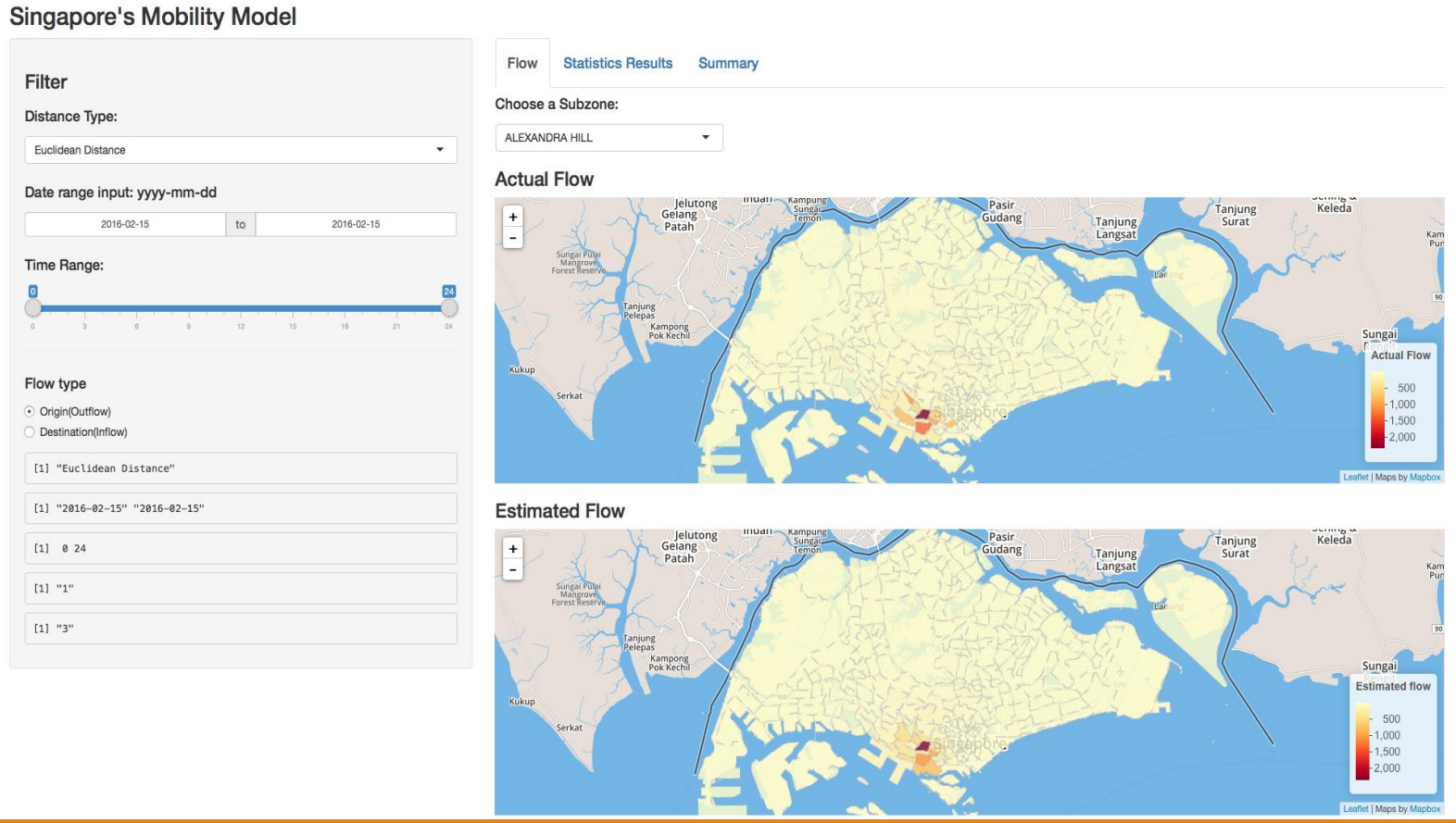
Application Framework

- **Technical Framework**



Demo

- **Model visualization Demo**



Singapore's Mobility Model

Filter

Distance Type:

Euclidean Distance

Date range input: yyyy-mm-dd

2016-02-15 to 2016-02-17

Time Range:

0 5 15 24

0 3 6 9 12 15 18 21 24

Flow type

- Origin(Outflow)
- Destination(Inflow)

```
[1] "Euclidean Distance"
```

```
[1] "2016-02-15" "2016-02-17"
```

```
[1]  5 15
```

```
[1] "2"
```

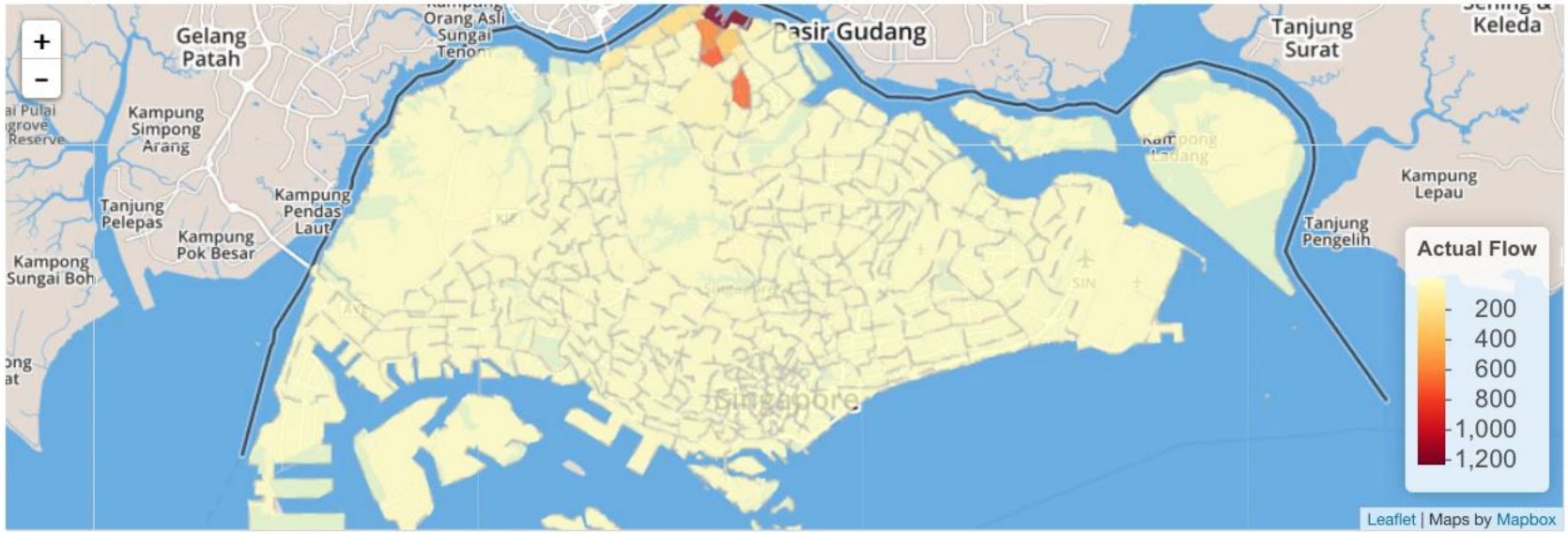
```
[1] "1"
```

Flow | Statistics Results | Summary

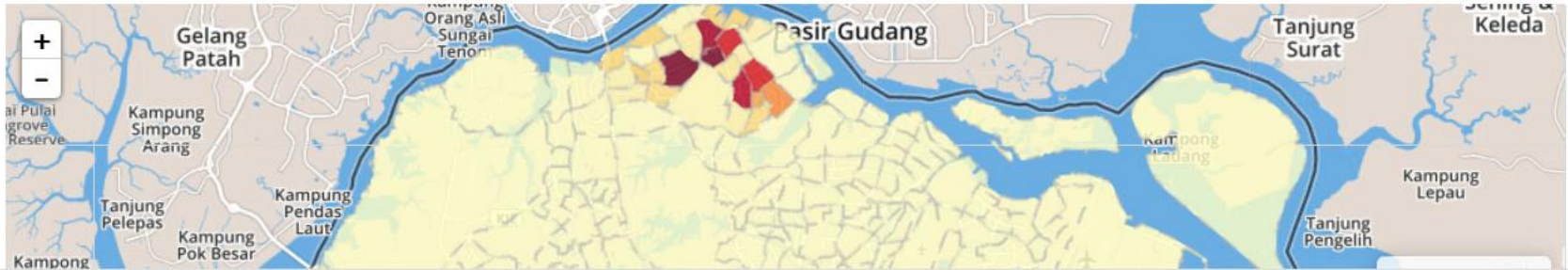
Choose a Subzone:

ADMIRALTY

Actual Flow



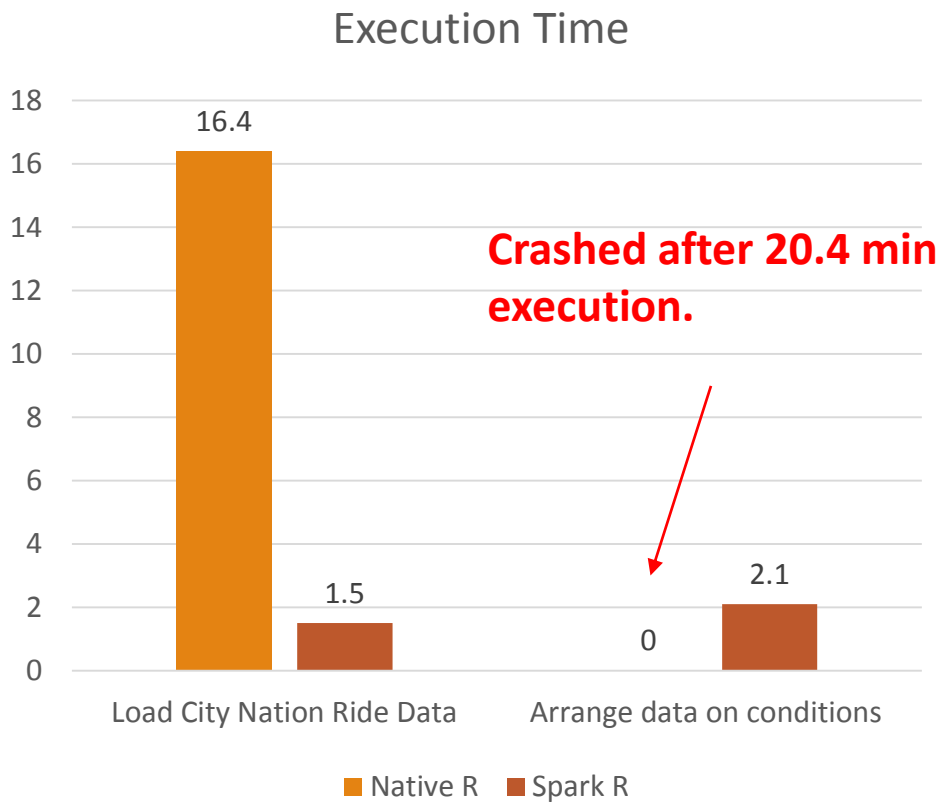
Estimated Flow



Analysis & Results

- **Data Preparation part**

Process	Data size	No Records	SparkR Time	Native R Time
Load City Nation Ride Data	4.82GB	50,697,539	1.5 min	16.4 min
Arrange data on conditions	4.82GB	50,697,539	2.1 min	NA



Analysis & Results

- **Modelling part**

Spark R Process	Data size	No Records
Run GLM	4.82GB	50,697,539
Run predict iteration loop	4.82GB	50,697,539

Native R Process	Data size	No Records
Run GLM	**16.1 MB (Data prepared by SparkR)**	104,329
Run predict iteration loop	**16.1 MB**	104,329

Conclusion & Future work

Conclusion

In conclusion, we investigate the possible approach to design and implement big data analytics application by integrating an open source big data processing framework and an open source data analysis environment.

Future Work

- We should introduce our project to other technical experts and test it with different scenarios.
- The flexibility of model calibration and results interface can be refined.

Thank you

Reference

- *Big data, 2014. http://en.wikipedia.org/wiki/Big_data , accessed Aug 2016.*
- *VITRIA. The Operational Intelligence Company, 2014. http://blog.vitria.com , accessed Aug, 2016.*
- *"Welcome to Apache Hadoop!". hadoop.apache.org. Retrieved 2015-12-16.*
- *"What is the Hadoop Distributed File System (HDFS)?". ibm.com. IBM. Retrieved 2014-10-30.*
- *Zaharia, Matei; Chowdhury, Mosharaf; Franklin, Michael J.; Shenker, Scott; Stoica, Ion. Spark: Cluster Computing with Working Sets USENIX Workshop on Hot Topics in Cloud Computing (HotCloud).*
- *SparkR: Scaling R Programs with Spark,2016.*
- *Apache Spark: http://spark.apache.org/*
- *Apache Hadoop: http://hadoop.apache.org/*
- *The R Project for Statistical Computing: https://www.r-project.org/*
- *Choosing R or Python for data analysis? An infographic, http://scl.io/Hi4twXG2#gs.BCyv5KM*
- *BIG DATA ANALYTICS – HADOOP PERFORMANCE ANALYSIS, Ketaki Subhash Raste, Spring 2014.*

Reference

- *AN EVALUATION OF THE SPARK PROGRAMMING MODEL FOR BIG DATA ANALYTICS, Haripriya Ayyalasomayajula May 2015.*
- *Big Data Analytics with Spark, http://www.virtualizationadmin.com/blogs/lowe/news/scale-up-vs-scale-out-the-key-differences-333.html*
- *Apache Spark a Big Data Analytics Platform for Smart Grid, Shyam Ra*, Bharathi Ganesh HBa, Sachin Kumar Sa, Prabaharan Poornachandranb, Soman K P, SMART GRID Technologies, August 6-8, 2015.*
- *Matrix Computations and Optimization in Apache Spark, Reza Bosagh Zadeh∗, Xiangrui Meng, Aaron Staple, January 1, 2016.*
- *Architectural Impact on Performance of In-memory Data Analytics: Apache Spark Case Study, Ahsan Javed Awan∗, Mats Brorsson∗, Vladimir Vlassov∗ and Eduard Ayguade, April 2016.*
- *WHITE PAPER: Using Visualization to Meet Big Data Challenges in Capital Markets, White Paper: Visualizing Big Data, www.Panopticon.com .*
- *Masterarbeit: Big Data and Machine Learning: A Case Study with Bump Boost, Maximilian Alber, Berlin, 19. Februar 2014.*

Reference

- *"The RedMonk Programming Language Rankings: June 2015 – tecosystems". Redmonk.com. 1 July 2015. Retrieved 10 September 2015.*
- *Kuhlman, Dave. "A Python Book: Beginning Python, Advanced Python, and Python Exercises")*
- *Big Data Analytics with Spark, Mohammed Guller,Apress.*
- *Discovering Spatial Patterns in Origin-Destination Mobility Data, Diansheng Guo, Department of Geography, University of South Carolina, 27 May 2012.*
- *Spatial interaction models, Wilson, 1971*
- *Estimating flows between geographical locations: 'get me started in' spatial interaction modelling, Adam Dennett Centre for Advanced Spatial Analysis, University College London.*
- *Agresti, A. (2002). Categorical data analysis: Wiley-Interscience.*
- *Batty, M., & Mackie, S. (1972). The calibration of gravity, entropy, and related models of spatial interaction. Environment and Planning, 4(2), 205-233.*
- *Bell, M. (2002). Comparing population mobility in Australia and New Zealand. Journal of Population Research, September (Special Issue), 169-193.*
- *Birkin, M., Clarke, G., & Clarke, M. (2010). Refining and Operationalizing Entropy- Maximizing Models for Business Applications. Geographical Analysis*

Reference

- *Entropy-Based Spatial Interaction Models for Trip Distribution, Geographical Analysis · October 2010*
- *SPATIAL INTERACTION MODELS OF INTERNATIONAL TELECOMMUNICATION FLOWS, Jean-Michel Guldmann, 40 Congress of the European Regional Science Association, August 29 - September 1, 2000.*
- *Shiny by RStudio http://shiny.rstudio.com/*
- *Leaflet for R https://rstudio.github.io/leaflet/*
- *DEX, Data city & Data nation, http://www.dex.sg/data/data-city-data-nation/*
- *Standard Error: http://onlinestatbook.com/2/regression/accuracy.html*
- *p-value: http://www.r-tutor.com/elementary-statistics/logistic-regression/significance-test-logistic-regression (P value for GLM R)*