CHƯƠNG 3: PHÂN TÍCH

3.1. Mô tả hệ thống

Hệ thống dữ liệu hỗ trợ nhà hoạch định Chính sách tăng cường năng lực giám sát, đánh giá và điều tiết vận tải một cách chủ động.

3.2. Đặc tả yêu cầu hệ thống

3.2.1. Yêu cầu chức năng:

#UC1: Tự động hóa pipeline dữ liệu:

- **Thu thập tự động:** Hệ thống có tác vụ được lập lịch tự động tải các file dữ liệu hành trình hàng tháng.
- **Biến đổi tự động:** Tự động thực hiện các quy trình để làm sạch, biến đổi, chuẩn hóa dữ liệu thô đã thu thập.
- **Nạp dữ liệu tự động:** Nạp dữ liệu tự động đã qua xử lý vào data warehouse.
- **Giám sát và cảnh báo:** Hệ thống có cơ chế ghi nhận lại trạng thái (thành công, thất bại) của mỗi lần chạy pipeline và gửi cảnh báo đến quản trị viên hệ thống trong trường hợp xảy ra lỗi.

#UC2: Cung cấp Dashboard KPIs:

a. Nội dung hiển thị:

- Tổng số chuyến đi
- Tổng doanh thu (total_amount)
- Tổng tiền tip (tip_amount)
- Tỷ lệ tiền tip trung bình (tip_amount / total_amount)
- Quãng đường trung bình mỗi chuyến (trip_distance)
- Thời gian trung bình mỗi chuyến (trip_duration_minutes)
- Giá cước trung bình mỗi chuyến (fare_amount)

b. Tương tác người dùng:

- **Bộ lọc thời gian:** chọn một khoảng thời gian (ví dụ từ ngày A đến ngày B), hoặc chọn nhanh theo năm, tháng.

- **Bộ lọc Loại dịch vụ:** Chọn một hoặc nhiều dịch vụ để so sánh (ví dụ: Yellow, Green, Uber, Lyft)

#UC3: Cung cấp Dashboard về nhu cầu di chuyển:

a. Nội dung hiển thị:

- Thành phần chính là bản đồ nhiệt của thành phố New York, tô đậm các vùng dựa trên mật độ (tổng lượt đón hoặc trả khách). Màu sắc càng đậm thể hiện mật độ càng cao.
- Xếp hạng 3 khu vực có lượt đón/trả khách cao nhất.

b. Tương tác người dùng

- **Bộ lọc theo thời gian:** Chọn một hoặc nhiều ngày trong tuần (ví dụ: thứ bảy, chủ nhật) và một hoặc nhiều khung giờ trong ngày (ví dụ: 07:00 - 9:00)
- **Bộ lọc theo loại dịch vụ:** Chọn một loại dịch vụ để xem mô hình hoạt động riêng của họ.
- **Bộ lọc theo Loại hành động:** Chuyển đổi bản đồ giữa chế độ xem "Lượt đón" và "Lượt trả".

#UC4: Cung cấp Dashboard về Giám sát mức độ hoạt động của các đơn vị

a. Nội dung hiển thị

- **Phân bố chuyến đi theo nhà cung cấp Công nghệ:** Hiển thị tổng số chuyến đi được ghi nhận (áp dụng cho Yellow và Green)
- **Phân bố Chuyến đi theo công ty HVFHS:** hiển thị tổng số chuyến đi thực hiện dưới mỗi taxi công nghệ (Urber, Lyft, Via)

b. *Tương tác người dùng*

- **Bộ lọc thời gian:** lọc sự thay đổi theo năm, quý, tháng dể theo dõi sự thay đổi về thị phần hoạt động của các đối tác qua thời gian.

#UC5: Cung cấp Dashboard về thu nhập của tài xế:

a. Nội dung thể hiện:

- Doanh thu trung bình mỗi giờ hoạt động có khách.
- Doanh thu trung bình trên mỗi dặm.

- Tỉ lệ đóng góp của tiền tip.

b. Tương tác người dùng:

- **Bộ lọc thời gian:** Lọc theo khoảng thời gian.
- **Bộ lọc Loại dịch vụ:** Chọn một loại dịch vụ để xem mô hình hoạt động riêng của họ.

3.2.2. Yêu cầu phi chức năng:

#UC1: Hiệu năng xử lý Pipeline:

Toàn bộ pipeline dữ liệu hàng tháng (từ lúc tải file tới khi nạp vào Data Warehouse) phải hoàn thành **dưới 60 phút.**

#UC2: Hiệu năng tải Dashboard:

Thời gian tải ban đầu của mỗi dashboard phải **dưới 15 giây**. Thời gian phản hồi sau mỗi lần tương tác với bộ lọc phải **dưới 5 giây**.

#UC3: Tính chính xác của dữ liệu:

Sai số giữa các chỉ số tổng hợp trên dashboard và kết quả tính toán thủ công trên một mẫu dữ liệu gốc phải **dưới 0.1%**.

***#UC4:* Khả năng mở rộng:**

Kiến trúc hệ thống phải có khả năng xử lý lượng dữ liệu tăng thêm 50% mà không cần thay đổi lớn về thiết kế.

***#UC5:* Tính sẵn sàng:**

Hệ thống (cụ thể là các dashboard) phải có độ sẵn sàng đạt 99.9% trong giờ hành chính.

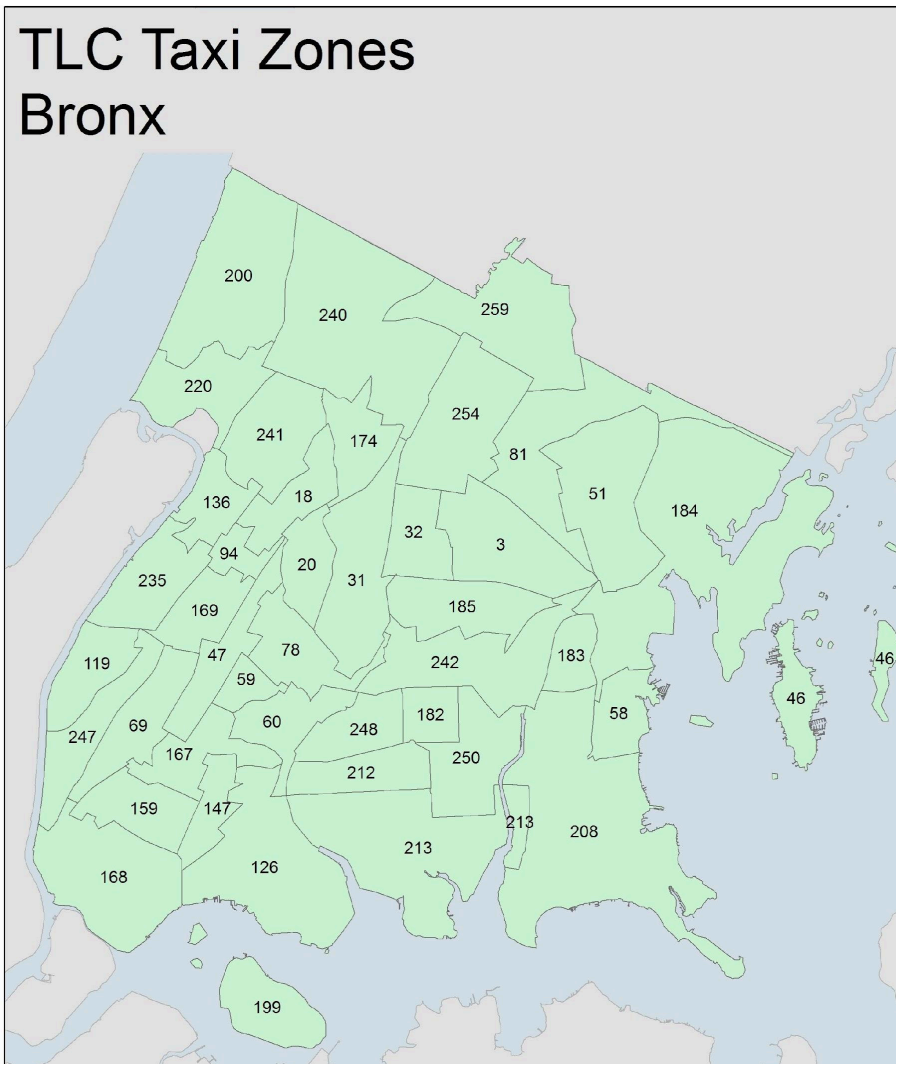
3.3. Phân tích nguồn dữ liệu:

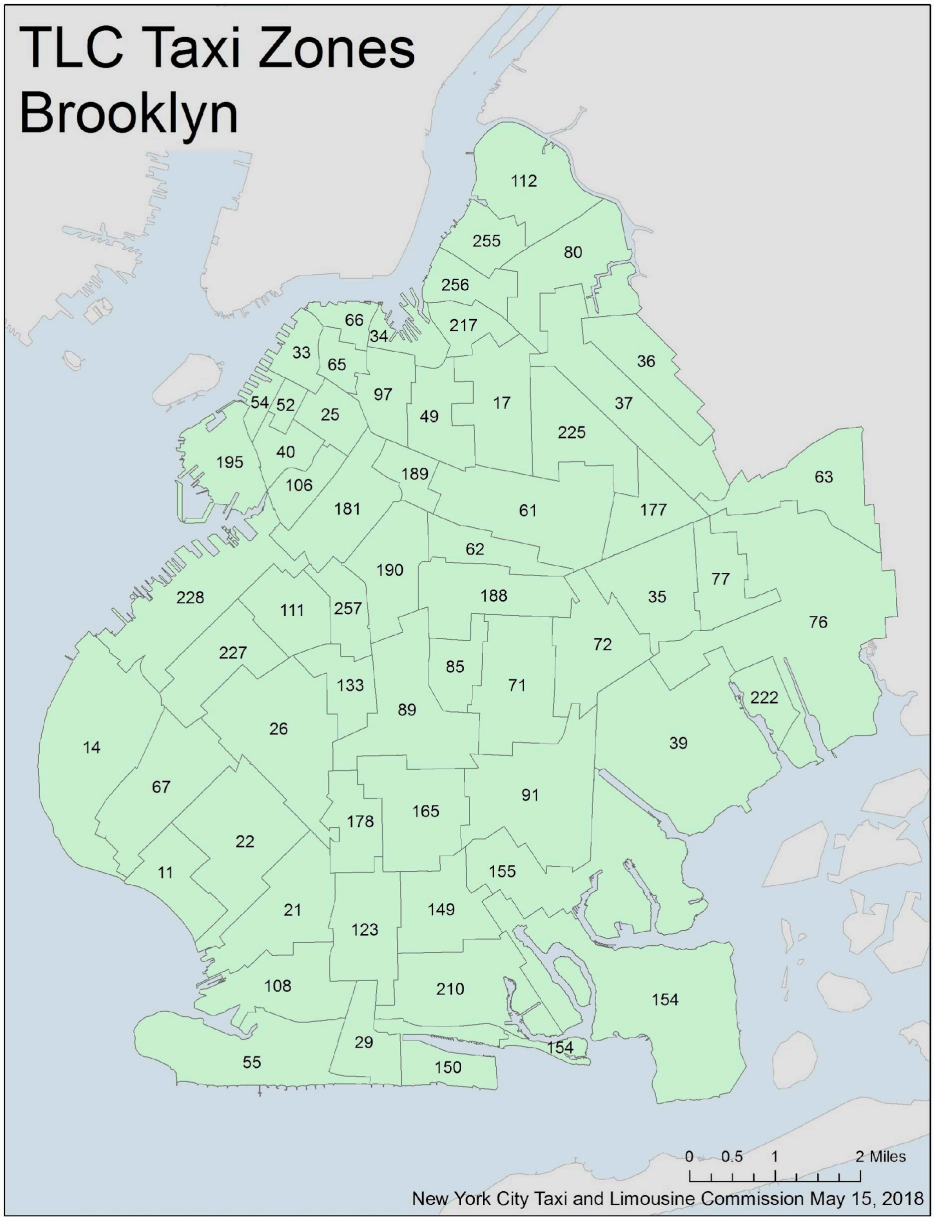
3.3.1. Nội dung

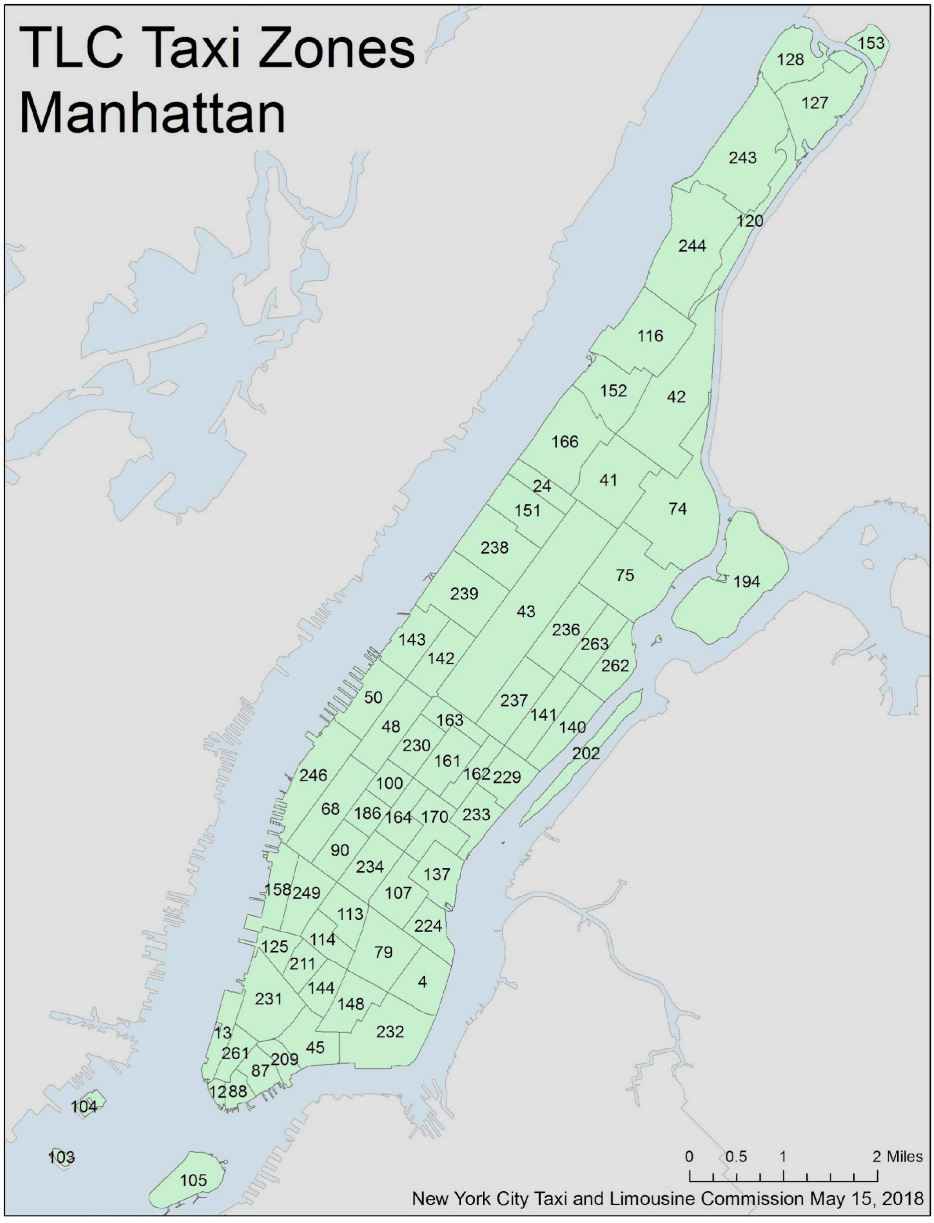
Bộ dữ liệu chứa thông tin về các chuyến đi được thực hiện bởi các phương tiện được cấp phép bởi TLC tại thành phố New York. Mỗi bảng ghi đại diện cho một chuyến đi duy nhất. Dữ liệu được chia làm ba loại chính:

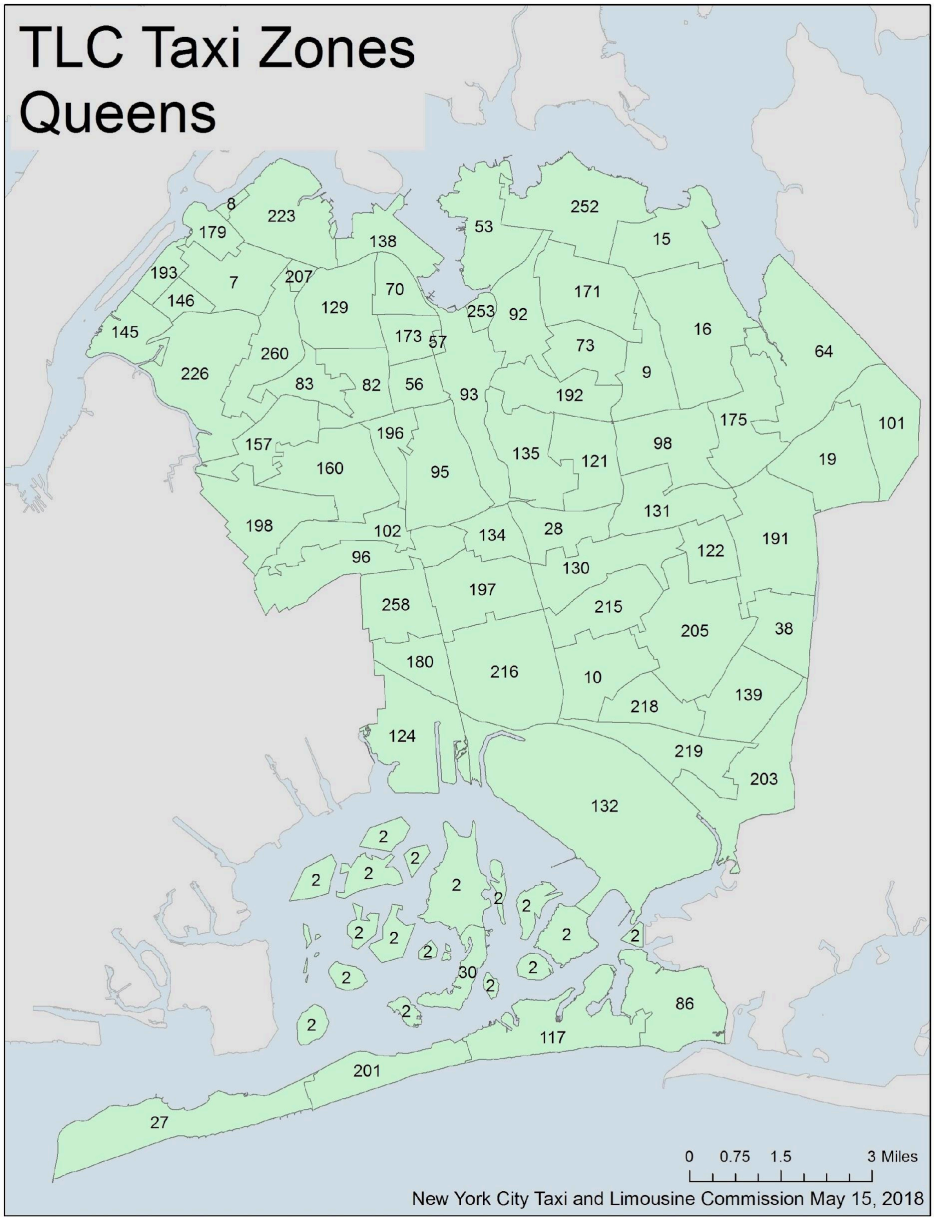
- Taxi vàng (Yellow Taxis): mang tính biểu tượng, được phép đón khách ở cả năm quận
- Taxi xanh (Green Taxis): dịch vụ taxi bên ngoài khu vực trung tâm Manhattan và chỉ được đón khách vẫy tay ở các khu vực được chỉ định.
- Phương tiện cho thuê (For-Hire Vehicles - FHV): bao gồm dữ liệu từ các dịch vụ gọi xe qua ứng dụng (Uber, Lyft, Via), xe Limousine sang trọng, và các loại xe dịch vụ.

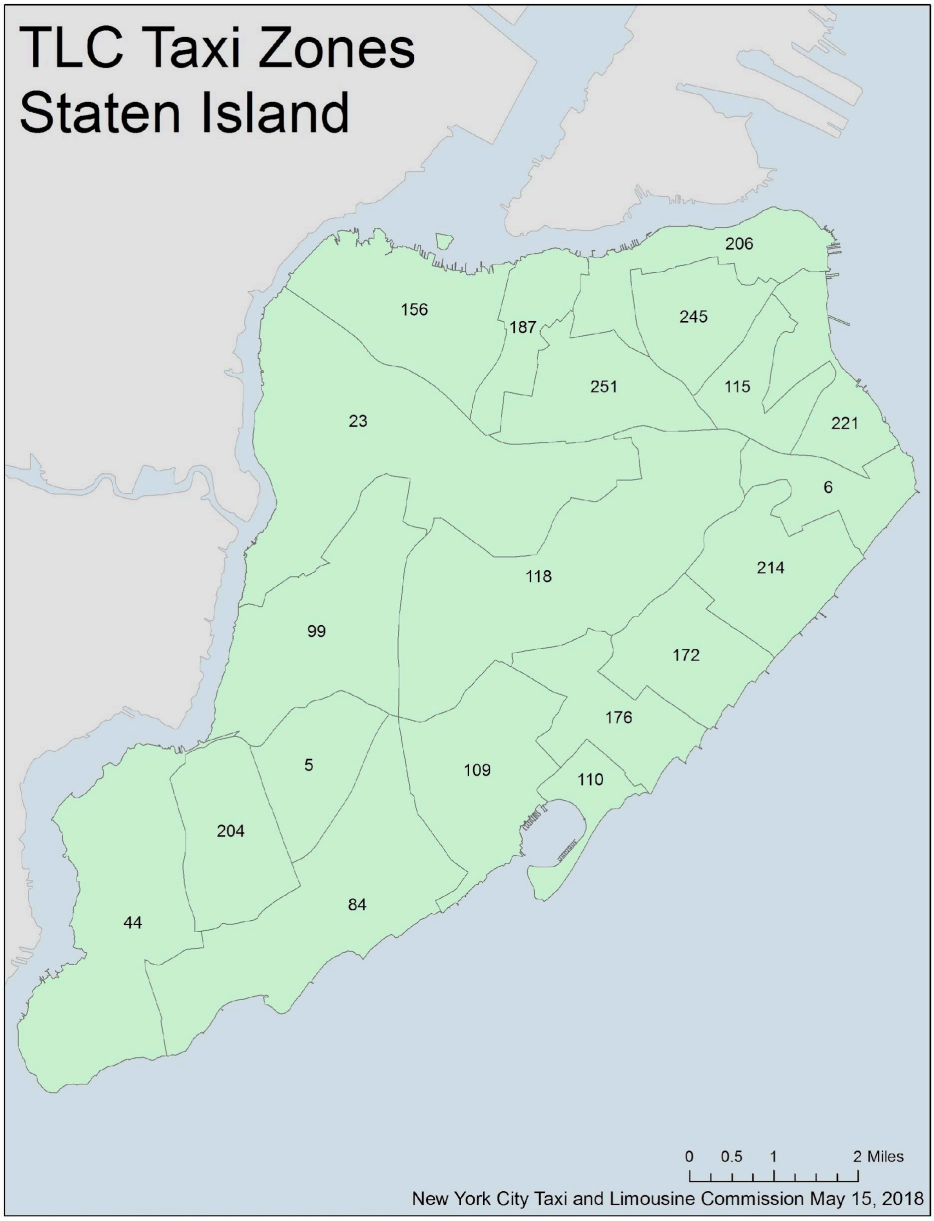
  Đây là thông tin bộ dữ liệu: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page


TLC Taxi Zones
Bronx
200
240
259
220
254
241
174
81
51
184
136
18
32
3
94
20
31
235
185
169
78
183
46
119
47
242
59
46
58
247
69
60
248
182
167
250
212
159
147
213
208
213
168
126
199


TLC Taxi Zones
Brooklyn
112
255
80
256
66
34
217
33
65
36
97
54
52
25
49
17
37
225
40
195
106
189
63
181
61
177
62
190
77
228
111
257
188
35
76
227
72
85
71
133
89
222
26
39
14
67
91
165
178
22
11
155
21
149
123
108
210
154
29
154
55
150
0 0.5 1 2 Miles
New York City Taxi and Limousine Commission May 15, 2018


TLC Taxi Zones
Manhattan
153
128
127
243
120
244
116
152
42
166
24
41
74
151
238
75
194
239
43
236
263
262
143
142
50
237
141
140
48
163
230
161
202
246
100
162
229
68
186
164
170
233
90
234
137
158
249
107
113
224
125
114
79
211
4
231
144
148
232
13
261
45
209
87
288
104
103
105
0 0.5 1 2 Miles
New York City Taxi and Limousine Commission May 15, 2018


TLC Taxi Zones
Queens
8
223
179
138
53
252
15
193
7
207
70
171
146
129
253
92
16
145
173
57
73
64
260
9
226
83
82
56
93
192
175
101
157
196
135
98
160
95
121
19
131
198
102
134
28
191
96
122
130
197
258
215
205
38
180
216
10
218
139
124
219
203
132
2
30
86
117
201
27
0
0.75
1.5
3 Miles
New York City Taxi and Limousine Commission May 15, 2018



3.3.2. Lịch sử bộ dữ liệu

Quá trình thu thập và công bố dữ liệu đã phát triển qua các năm:

- 2009: TLC bắt đầu nhận dữ liệu chuyến đi từ các nhà cung cấp công nghệ cho taxi.
- 2013: Taxi xanh được bổ sung vào hệ thống.
- 2015:

+ Dữ liệu của Taxi Vàng và Xanh được công bố công khai lần đầu tiên trên cổng thông tin Open Data.
+ TLC bắt đầu nhận dữ liệu chuyến đi từ tất cả các cơ sở FHV. Ban đầu, dữ liệu này còn hạn chế, chỉ bao gồm mã cơ sở điều phối, thời gian và địa điểm đón khách.
- 2016: Dữ liệu FHV bắt đầu được công bố rộng rãi.
- 2017: Yêu cầu báo cáo cho FHV được mở rộng, bao gồm thêm thời gian và địa điểm trả khách. Thông tin về các chuyến đi chia sẻ (shared rides) cũng bắt đầu được thu thập.
- 2019: TLC tạo ra một hạng giấy phép mới là "Dịch vụ Cho thuê Khối lượng lớn" (High Volume For-Hire Services - HVFHS) cho các công ty có trên 10.000 chuyến đi mỗi ngày.
- 2025: Một cột mới cdb_congestion_fee được thêm vào dữ liệu của taxi vàng, xanh và HVFHS.

3.3.3. Độ lớn bộ dữ liệu và định dạng của dữ liệu

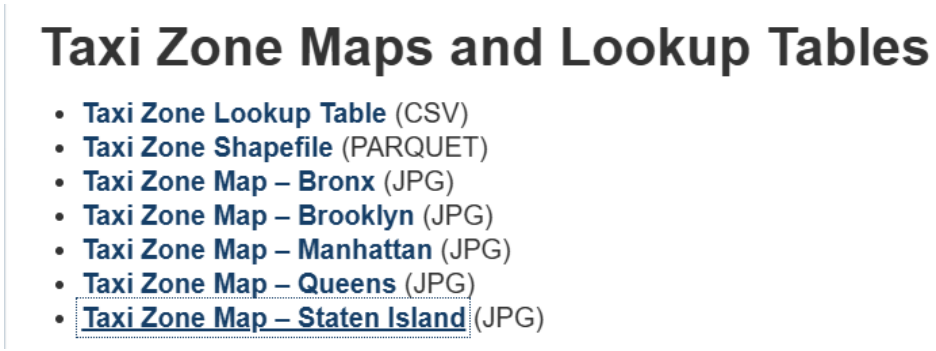
3.3.3.1. Độ lớn dữ liệu

Dữ liệu có quy mô rất lớn. Một tháng dữ liệu có thể **hơn một triệu bản ghi.**

3.3.3.2. Định dạng dữ liệu

- File dữ liệu các năm được cung cấp ở định dạng Parquet. Là định dạng lưu trữ dạng cột (column storage), giảm kích thước tệp dữ liệu và tăng tốc độ đọc dữ liệu đáng kể.

- File dữ liệu tham chiếu Lookup table ở dạng file CSV.



3.3.4. Cấu trúc dữ liệu (data schema)

Nguồn dữ liệu bao gồm: TLC Trip Record, Taxi Zone Lookup

Bảng dữ liệu taxi vàng

Tên thuộc tính	Mô tả
VendorID	Mã của nhà cung cấp công nghệ 1 = Creative Mobile Technologies, LLC 2 = Curb Mobility, LLC 6 = Myle Technologies Inc 7 = Helix
tpep_pickup_datetime	Ngày và giờ đồng hồ tính cước được bật
tpep_dropoff_datetime	Ngày và giờ đồng hồ tính cước được tắt.
passenger_count	Số lượng hành khách trên xe
trip_distance	Quãng đường của chuyến đi (dặm) do đồng hồ đo được.
RatecodeID	Mã loại giá cước cuối cùng: 1 = Standard rate (Tiêu chuẩn): chuyến đi thông thường dựa trên đồng hồ đo; 2 = JFK và 3 = Newark: các chuyến đi đến hoặc đi từ hai sân bay lớn John F. Kennedy (JFK) và Newark (EWR). Áp dụng mức giá cước cố định; 4 = Nassau or Westchester: Các chuyến đi đến các quận ngoại ô Nassau hoặc Westchester, nằm ngoài địa phận thành phố New York;

	5 = Negotiated fare (giá thương lượng): Khách hàng và tài xế đã thương lượng và đồng ý với một mức giá; 6 = Group ride: chuyến đi tài xế đón nhiều khách hàng khác nhau trên cùng một chuyến xe; 99 = Null/unknown - không xác định;
store_and_fwd_flag	Cờ báo hiệu chuyến đi có được lưu vào bộ nhớ xe khi trước khi gửi đi hay không (do mất kết nối) Y: được lưu trữ. N: chưa được lưu trữ.
PULocationID	ID khu vực taxi của TLC nơi bắt đầu chuyến đi.
DULocationID	ID khu vực taxi của TLC nơi kết thúc chuyến đi.
payment_type	Mã phương thức thanh toán 1=Credit card, 2=Cash, 3=No charge, 4=Dispute, 5=Unknown, 6=Voided.
fare_amount	Cước phí theo thời gian và quãng đường
extra	Các khoản phụ phí và phụ thu khác.
mta_tax	Thuế MTA được kích hoạt tự động
tip_amount	Tiền boa (chỉ áp dụng cho thanh toán thẻ)

tolls_amount	Phí cầu đường
improvement_surcharge	Phụ phí cải thiện
total_amount	Tổng tiền thanh toán
congestion_surcharge	Phụ phí tắc nghẽn
airport_fee	Phí sân bay, chỉ áp dụng cho các chuyến đi đón tại LaGuardia và JFK.
cbd_congestion_fee	Phí tắc nghẽn khu vực trung tâm (từ 2025).

Bảng dữ liệu taxi xanh

Tên thuộc tính	Mô tả
VendorID	Mã của nhà cung cấp dịch vụ công nghệ (LPEP)
lpep_pickup_datetime	Ngày và giờ đồng hồ tính cước được bật.
lpep_dropoff_datatime	Ngày và giờ đồng hồ tính cước được tắt.
store_and_fwd_flag	Cờ báo hiệu chuyến đi có được lưu vào bộ nhớ xe trước khi gửi đi hay không ?
RatecodeID	Mã loại giá cước cuối cùng 1=Standard, 2=JFK,

	3=Newark, 4=Nassau/Westchester, 5=Negotiated, 6=Group ride
PULocationID	ID khu vực taxi của TLC nơi bắt đầu chuyến đi
DOLocationID	ID khu vực taxi của TLC nơi kết thúc chuyến đi
passenger_count	Số lệnh hành khách trên xe
trip_distance	Quãng đường của chuyến đi do đồng hồ đo được.
fare_amount	Cước phí tính theo thời gian và quãng đường.
extra	Các khoản phụ phí và phụ thu khác.
mta_tax	Thuế MTA được kích hoạt tự động
tip_amount	Tiền boa (chỉ áp dụng cho thanh toán thẻ)
tolls_amount	Phí cầu đường
improvement_surcharge	Phụ phí cải thiện
total_amount	Tổng tiền thanh toán
payment_type	Mã phương thức thanh toán 1=Credit card, 2=Cash, 3=No charge, 4=Dispute, 5=Unknown, 6=Voided.

trip_type	Mã cho biết loại chuyến đi: 1=Street-hail 2=Dispatch
congestion_surcharge	Phụ phí tắc nghẽn của bang New York.
cbd_congestion_fee	Phí cho Khu vực giảm tắc nghẽn của MTA, bắt đầu từ 05/01/2025.

Bảng dữ liệu Phương tiện cho thuê - khối lượng lớn (High-volume FHVs)

Tên thuộc tính	Mô tả
hvfhs_license_num	Số giấy phép TLC của cơ sở hoặc doanh nghiệp HVFHS. Tính đến tháng 9 năm 2019, các bên được cấp phép HVFHS bao gồm: HV0002: Juno HV0003: Uber HV0004: Via HV0005: Lyft
dispatching_base_num	Số Giấy phép Cơ sở của TLC của cơ sở đã điều phối chuyến đi.
originating_base_num	Số hiệu cơ sở của cơ sở đã nhận yêu cầu chuyến đi ban đầu.
request_datetime	Ngày giờ khách hàng yêu cầu được đón.

on_scene_datetime	Ngày giờ tài xế đến địa điểm đón (chỉ dành cho xe hỗ trợ người khuyết tật)
pickup_datetime	Ngày và giờ đón khách của chuyến đi.
dropoff_datetime	Ngày và giờ trả khách của chuyến đi.
PULocationID	Khu vực taxi của TLC nơi chuyến đi bắt đầu.
DOLocationID	Khu vực taxi của TLC nơi chuyến đi kết thúc.
trip_miles	Tổng số dặm cho chuyến đi của hành khách.
trip_time	Tổng thời gian tính bằng giây cho chuyến đi của hành khách.
base_passenger_fare	Giá cước cơ bản của hành khách trước phí cầu đường, tiền boa, thuế và các loại phí khác.
tolls	Tổng số tiền phí cầu đường đã trả trong chuyến đi.
bcf	Tổng số tiền thu được trong chuyến đi cho Quỹ Black Car.
sales_tax	Tổng số tiền thu được trong chuyến đi cho thuế bán hàng của bang New York.
congestion_surcharge	Tổng số tiền thu được trong chuyến đi cho phụ phí tắc nghẽn của bang New York.
airport_fee	$2.50 cho cả việc trả và đón khách tại các sân bay LaGuardia, Newark và John F. Kennedy.
tips	Tổng số tiền boa nhận được từ hành khách.

driver_pay	Tổng lương tài xế (không bao gồm phí cầu đường hoặc tiền boa và sau khi trừ hoa hồng, phụ phí hoặc thuế).
shared_request_flag	Hành khách có đồng ý đi chung xe/đi ghép hay không, bất kể họ có được ghép cặp hay không? (Y/N)
shared_match_flag	Hành khách có đi chung xe với một hành khách khác đã đặt xe riêng vào bất kỳ thời điểm nào trong chuyến đi không? (Y/N)
access_a_ride_flag	Chuyến đi có được thực hiện thay mặt cho Cơ quan Giao thông Vận tải Đô thị (MTA) không? (Y/N)
wav_request_flag	Hành khách có yêu cầu một chiếc xe cho người khuyết tật (WAV) không? (Y/N)
wav_match_flag	Chuyến đi có diễn ra trên một chiếc xe cho người khuyết tật (WAV) không? (Y/N)
cbd_congestion_fee	Phí mỗi chuyến cho Khu vực Giảm tắc nghẽn của MTA bắt đầu từ ngày 5 tháng 1 năm 2025.

Bảng dữ liệu Bảng dữ liệu Phương tiện cho thuê khác (FHV)

Tên thuộc tính	Mô tả
dispatching_base_num	Mã giấy phép của cơ sở đã điều phối chuyến đi.
pickup_datetime	Ngày và giờ đón khách.
dropOff_datetime	Ngày và giờ trả khách.

PULocationID	ID Khu vực Taxi của TLC nơi bắt đầu chuyến đi.
DOlocationID	ID Khu vực Taxi của TLC nơi kết thúc chuyến đi.
SR_Flag	Cờ báo hiệu chuyến đi chia sẻ. Lưu ý có sự khác biệt trong cách Lyft gắn cờ này so với các công ty khác.
Affiliated_base_num	Mã giấy phép của cơ sở mà phương tiện liên kết.

Bảng Tra cứu Khu vực Taxi (Taxi Zone Lookup Table)

Tên thuộc tính	Mô tả
LocationID	Khóa chính. ID của khu vực (từ 1-263), dùng để nối với PULocationID và DOLocationID.
Borough	Tên quận (ví dụ: Manhattan, Brooklyn, Queens).
Zone	Tên cụ thể của khu vực (ví dụ: JFK Airport, Upper East Side South).
service_zone	Vùng dịch vụ mà khu vực đó thuộc về (ví dụ: Boro Zone, Yellow Zone, EWR).

Bảng Ánh xạ Cơ sở Điều phối HVFHS (HVFHS Base Mapping)

Tên Thuộc tính	**Mô tả**
hvfhs_license_num	Mã số giấy phép dịch vụ khối lượng lớn (ví dụ: 'HV0003').

dispatching_base_num	Khóa chính. Mã số của cơ sở điều phối, dùng để nối với trường cùng tên trong dữ liệu FHV/HVFHS.
base_name	Tên pháp lý của cơ sở điều phối.
app_company_affiliation	Tên công ty mẹ (ví dụ: 'Uber', 'Lyft', 'Juno', 'Via').

3.3.5. Đối chiếu dữ liệu với Yêu cầu:

Bảng Đối chiếu chi tiết cho UC2: Dashboard "Giám sát KPIs"

Yêu cầu Chỉ số (KPI)	**Thuộc tính Dữ liệu Nguồn Cần thiết**	**Ghi chú**	**Độ bao phủ**
Tổng số chuyến đi	(Không cần trường cụ thể, đếm số dòng)	**Khả thi.** Mỗi dòng trong file dữ liệu đại diện cho một chuyến đi.	Yellow taxi, Green taxi, HVFHS, FHV
Tổng doanh thu	total_amount	**Khả thi.** Cần xử lý các giá trị âm hoặc bằng 0.	Chỉ có ở yellow taxi, Green taxi, HVFHS. Hoàn toàn thiếu ở FHV thông thường.
Tổng tiền tip	tip_amount	**Khả thi.** Trường này có sẵn. Cần xử lý các giá trị âm.	Chỉ có ở Yellow, Green, HVFHS.
Quãng đường trung bình	trip_distance	**Khả thi.** Trường này có sẵn. Cần xử lý các giá trị âm để tránh sai	Chỉ có ở Yellow, Green, HVFHS. Thường thiếu ở FHV cũ.

		lệch kết quả trung bình.	
Thời gian trung bình	tpep_pickup_datetime, tpep_dropoff_datetime	**Khả thi, cần tính toán.** Thuộc tính này không có sẵn, phải được **suy ra** bằng cách tính hiệu số giữa hai trường thời gian. Cần xử lý trường hợp dropoff_datetime trước pickup_datetime.	Tất cả các file đều có pickup_datetime. dropoff_datetime **bị thiếu** ở các file FHV cũ trước 2017.
Bộ lọc Loại dịch vụ	service_type	**Khả thi, cần tạo mới.** Dữ liệu gốc không có trường service_type. Thuộc tính này phải được **tạo ra** trong quá trình xử lý dựa trên nguồn gốc của file (ví dụ: file yellow_... -> service_type = 'Yellow').	Thuộc tính không có sẵn nhưng có thể tạo ra một cách logic cho tất cả các bảng ghi.

Các thuộc tính quan trọng nhất là **total_amount, trip_distance,** và **các trường datetime.** Cần tập trung vào việc **tính toán trip_duration** và **tạo mới trường service_type.**

Bảng Đối chiếu chi tiết cho UC3: Dashboard "Nhu cầu di chuyển”

Yêu cầu chức năng	**Thuộc tính dữ liệu nguồn cần thiết**	**Ghi chú**	**Độ bao phủ**
Hiển thị mật độ theo khu vực	PULocationID, DOLocationID	Khả thi. Cần join với file Taxi Zones Lookup để có được tên Quận (Borough) và khu vực (Zones)	Có ở tất cả các file. Cần join với bảng tra cứu để làm hiển thị thêm thông tin.
Bộ lọc Giờ trong Ngày	tpep_pickup_datetime	Khả thi. Cần trích xuất thành phần giờ từ trường datatime để phục vụ lọc.	Có ở tất cả các file. Chú ý các bản ghi có datatime không hợp lệ.
Bộ lọc Giờ trong Ngày	tpep_pickup_datetime	Khả thi. Cần trích xuất thành phần “ngày trong tuần” từ trường datetime.	Có ở tất cả các file. Chú ý các bản ghi có datatime không hợp lệ.

Các thuộc tính quan trọng nhất là **PULocationID** và **DOLocationID**. Nhưng cần kết hợp với bảng Taxi Zones Lockup.

Bảng Đối chiếu chi tiết cho UC4: Dashboard "Giám sát mức độ hoạt động của các đơn vị”

Yêu cầu chỉ số	**Thuộc tính dữ liệu nguồn cần thiết**	**Ghi chú**	**Độ bao phủ**
Phân bổ theo Nhà cung cấp Công nghệ	VendorID	Khả thi. Dùng đo lường hoạt động của các công ty công nghệ.	Chỉ ở Yellow và Green taxi.
Phân bổ theo công ty HVFHS	hvfhs_license_num	Khả thi. Dùng đo lường hoạt động	Chỉ ở bộ dữ liệu HVFHS

		của các công ty vận tải lớn như Uber, Lyft.	

Mang tính cục bộ cao cho từng loại hình vận tải trên thị trường. Dashboard được thiết kế riêng biệt cho từng loại hình vận tải, giúp TLC có cái nhìn chi tiết về từng phân khúc trong hệ sinh thái xe vận tải.

Bảng Đối chiếu chi tiết cho UC5: Dashboard "Thu nhập của tài xế"

Yêu cầu chỉ số	**Thuộc tính dữ liệu nguồn cần thiết**	**Ghi chú**	**Độ bao phủ**
Tính doanh thu / giờ hoạt động	Total_amount, base_passenger_fare, tips, tolls, pickup datatime, dropoff datetime.	Khả thi, **total_amount** cần được tính toán nhất quán từ các thành phần phí khác nhau giữa các dịch vụ. **trip_duration** phải được suy ra chính xác.	Chỉ ở Yellow, Green và HVFHS.
Tính doanh thu / dặm	trip_distance, trip_miles	Khả thi. Tương tự như trên, cần chuẩn hóa trip_distance và trip_miles thành một đơn vị chung.	Chỉ áp dụng cho các dịch vụ có dữ liệu về quãng đường.
Tính Tỷ lệ Tiền tip	tip_amount, tips, total_amount	Khả thi. Cần một công thức tính total_amount nhất quán để làm mẫu số.	Chỉ áp dụng cho các dịch vụ có dữ liệu về tiền tip.

Cần xây dựng **logic nghiệp vụ vững** để chuẩn hóa và tính toán các chỉ số tài chính.

Bảng Tổng hợp các Thuộc tính Dữ liệu Quan trọng và Chiến lược Xử lý

Thuộc tính Cốt lõi	**Vai trò**	**Thách thức chính & Chiến lược Xử lý**
LocationIDs và datetimes	Nền tảng cho các phân tích toàn thị trường (địa lý, xu hướng).	**Thách thức:** Tên cột không nhất quán. **Chiến lược:** Chuẩn hóa tên cột, làm giàu dữ liệu từ bảng tra cứu, xử lý logic thời gian sai.
Thuộc tính Tài chính & Hiệu suất	Nền tảng cho các phân tích KPI.	**Thách thức:** Cục bộ, chỉ có ở một vài dịch vụ. **Chiến lược:** Xử lý có điều kiện, chỉ áp dụng phép tính cho các dịch vụ có dữ liệu và thiết kế dashboard để thể hiện rõ điều này.
service_type	Chiều phân tích chính để so sánh các phân khúc.	**Thách thức:** Không có sẵn. **Chiến lược:** Feature Engineering - tạo ra một cột thống nhất dựa trên nguồn gốc file.
Thuộc tính Định danh Đối tác	Nền tảng cho phân tích hoạt động của các công ty.	**Thách thức:** Không nhất quán về ý nghĩa. **Chiến lược:** Điều chỉnh phạm vi yêu cầu, phân tích riêng cho từng loại định danh (Vendor, hvfhs_license_num).