**CSC12107 – HTTT PHỤC VỤ TRÍ TUỆ KINH DOANH**

**ĐỒ ÁN MÔN HỌC**

**Đề tài: Xây dựng và phân tích Dữ liệu về lộ trình Yellow Taxi ở New York City (NYC) từ năm 2014 đến 2018.**

**Báo cáo lần 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Mã nhóm*** | ***MSSV*** | ***Họ và tên*** | **SĐT** | ***Ghi chú*** |
| **CQ-BI-24** | **1712904** | **Nguyễn Bách Việt** | 0937150768 |  |
| **1712923** | **Phạm Lê Hoài Vũ** | 0368804297 |  |
| **1712929** | **Nguyễn Phượng Vỹ** | 0385092206 | **Nhóm trưởng** |

**MỤC LỤC**

[project description 4](#_Toc56771055)

[1. About Project: 4](#_Toc56771056)

[2. Key Deliverables 5](#_Toc56771057)

[3. Critical Success Factors 5](#_Toc56771058)

[4. Risks and Concerns 5](#_Toc56771059)

[BI Dimensional Logical Model Design 5](#_Toc56771060)

[1. DATA STORE: YELLOW TAXI 2014 5](#_Toc56771061)

[2. DATASOURCE ANALYSIS 6](#_Toc56771062)

[**2.1** **Yellow Taxi 2014** 6](#_Toc56771063)

[**a)** **Mô tả dữ liệu:** 6](#_Toc56771064)

[**b)** **Nguồn dữ liệu:** 6](#_Toc56771065)

[**c)** **Hình ảnh nguồn dữ liệu** 6](#_Toc56771066)

[**d)** **Bảng phân tích nguồn dữ liệu** 7](#_Toc56771067)

[**2.2** **New York City Census Tracts for 2010 US Census** 9](#_Toc56771068)

[**a)** **Mô tả dữ liệu** 9](#_Toc56771069)

[**b)** **Nguồn dữ liệu** 9](#_Toc56771070)

[**c)** **Hình ảnh nguồn dữ liệu** 9](#_Toc56771071)

[**d)** **Bảng phân tích nguồn dữ liệu** 10](#_Toc56771072)

[**e)** **Giải thích một số thuộc tính** 12](#_Toc56771073)

[3. YÊU CẦU CẦN SỬ DỤNG API 15](#_Toc56771074)

[**3.1** **Từ tọa độ xác định vị trí, số nhà, đường, quận, thành phố, bang, nước.** 15](#_Toc56771075)

[**a)** **API sử dụng:** 15](#_Toc56771076)

[**b)** **Bảng phân tích dữ liệu trả về** 16](#_Toc56771077)

[**c)** **Quá trình tổng hợp dữ liệu** 17](#_Toc56771078)

[**3.2** **Từ tọa độ xác định Census Block ID (FIPS code)** 17](#_Toc56771079)

[**a)** **API sử dụng:** 18](#_Toc56771080)

[**b)** **Bảng phân tích dữ liệu trả về** 19](#_Toc56771081)

[**c)** **Quá trình tổng hợp dữ liệu** 19](#_Toc56771082)

[4. KIẾN TRÚC DATA MART BÊN TRONG 20](#_Toc56771083)

[***REQ01****:* Người dùng nghiệp vụ cần phân tích dữ liệu các chuyến xe đã được thực hiện vào những thời gian xác định theo khu vực địa lý. Cũng như sự phân bố lượng đặt xe được ở các vùng 20](#_Toc56771084)

[**1.** **Phân tích yêu cầu** 20](#_Toc56771085)

[**2.** **Phân tích yêu cầu về dữ liệu** 20](#_Toc56771086)

[**3.** **Xác định dimension table + SCD:** 22](#_Toc56771087)

[***REQ02****:* Người dùng nghiệp vụ cần phân tích về việc kinh doanh từ dịch vụ taxi. Cần biết doanh thu, chi phí, lợi nhuận theo năm, kỳ, quý, tháng, ngày cũng như loại thanh toán (đơn vị tiền tệ là dollar) 23](#_Toc56771088)

[**1.** .**Phân tích yêu cầu** 23](#_Toc56771089)

[**2.** **Phân tích yêu cầu về dữ liệu.** 24](#_Toc56771090)

[**3.** **Xác định Dimension table + SCD:** 26](#_Toc56771091)

[Business Process 29](#_Toc56771092)

[1. Quy trình đặt xe Taxi 29](#_Toc56771093)

[2. Quy trình điều động xe Taxi 29](#_Toc56771094)

[Thiết kế data flow 30](#_Toc56771095)

[I. Hiện trạng hệ thống 30](#_Toc56771096)

[II. Xác định nhu cầu 30](#_Toc56771097)

[III. Yêu cầu về chức năng 30](#_Toc56771098)

[IV. Yêu cầu về dữ liệu 31](#_Toc56771099)

[V. Yêu cầu về hệ thống 31](#_Toc56771100)

[VI. Kiến trúc luồng dữ liệu 32](#_Toc56771101)

[Thiết kế KIẾN TRÚC DỮ LIỆU 32](#_Toc56771102)

[Approvals 33](#_Toc56771103)

[Vấn đáp - zoom 33](#_Toc56771104)

# project description

Xây dựng và phân tích Dữ liệu về lộ trình Yellow Taxi ở New York City (NYC) từ năm

2014 đến 2018.

## About Project:

1. **Mô tả dữ liệu:**

* Dữ liệu Yellow Taxi Trip năm 2014:
* Dữ liệu Yellow Taxi Trip từ năm 2014 đến 2018 (https://data.cityofnewyork.us)
* Dữ liệu Census Block (download Geospatial Data, ví dụ GeojSON):

https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku

1. **Yêu cầu:**

* Mô tả ý nghĩa các thuộc tính của 2 nguồn dữ liệu trên
* Thiết kế kho dữ liệu (KDL) và tổng hợp, nạp dữ liệu các nguồn vào KDL
* Sử dụng dịch vụ Google API Reverse Geocoding (hoặc tương tự) để lấy địa chỉ (street, district, city, state) từ tọa độ đón và trả khách
* Sử dụng dịch vụ hoặc các thư viện có sẵn để lấy Census Block ID từ tọa độ đón và trả khách
* Thiết kế và xây dựng Cube

**Gợi ý:**

* Chuyển đổi dữ liệu ngày tháng sao cho có thể tạo được Date dimension hierachy Year-Quarter-Month-Date trong Cube
* Xác định và thiết kế các phân cấp chiều còn lại
* **Khai thác dữ liệu:**
* **OLAP:**

Phân tích chuyến đi theo Geography (street, district, city, state, boroughs) ..., theo Census Block, theo thời gian, theo loại thanh toán...

* **Report:**
* Dùng regional map để biểu diễn trực quan (bằng màu sắc) sự phân bố số lượng đón xe (pickups) và số lượng trả khách (drop-offs) ở các vùng (district, city) theo thời gian tháng, quý, năm?
* Dùng regional map để biểu diễn trực quan (bằng màu sắc) sự phân bố số lượng đón xe (pickups) và số lượng trả khách (drop-offs) ở các Census Block theo thời gian tháng, quý, năm?
* Phân tích thời điểm nào trong ngày là giờ cao điểm và thấp điểm của các chuyến taxi? Vẽ đồ thị phân bố số lượng đón taxi theo giờ trong ngày trong tất cả các năm.
* Thống kê doanh thu năm theo quận (district) đón khách.
* Dưới góc độ là người khai thác dữ liệu, SV tự đề xuất và thực hiện các nhu cầu phân tích khác.
* **Prediction**

Gợi ý:

* Dự đoán vào một thời điểm cụ thể trong năm theo từng district thì lượng khách đón taxi sẽ như thế nào
* Để phân chia thành phố New York thành khu vực để có thể thực hiện dự đoán theo khu vực 🡪 sử dụng thuật toán K-mean...
* Xây dựng một mô hình dự đoán cho số tiền boa theo tỷ lệ phần trăm của tổng giá vé
* Sinh viên tự đề xuất 1 yêu cầu phân tích, lựa chọn mô hình phù hợp.

1. **Kết luận chung**

## Key Deliverables

**G3.3** Thiết kế lược đồ chuẩn hoá, đa chiều (sao, bông tuyết) dựa vào dữ liệu hệ thống tác vụ và yêu cầu phân tích từ tình huống cho trước Mô tả bài tập

**G5.1** Triển khai quy trình ETL để rút trích dữ liệu từ nhiều nguồn, biến đổi, làm sạch dữ liệu, nạp dữ liệu vào KDL sử dụng công cụ SSIS

**G5.2** Xây dựng KDL đa chiều sử dụng công cụ SSAS và giải thích được lựa chọn phép toán OLAP phù hợp đối với 1 số yêu cầu phân tích.

**G5.3** Sử dụng một số công cụ biểu diễn dữ liệu (SSRS, powerBI, excel...) để biểu diễn kết quả phân tích, khai thác được (report, dashboard...)

**G5.4** Sử dụng SSAS và áp dụng các kỹ thuật mining tích hợp để thực hiện khai thác dữ liệu từ KDL xây dựng được.

## Critical Success Factors

* ETL process (data flow, data cleaning, ETL data from source to DW)
* Project hoàn chỉnh (khai thác KDL với report, olap, mining, tạo job tự động định kỳ thực hiện ETL)

## Risks and Concerns

* Dữ liệu lớn, có thể sai sót thông tin.
* Có thể không ứng dụng cho thực tế nhiều.

# BI Dimensional Logical Model Design

## DATA STORE: YELLOW TAXI 2014

Data Store: Yellow Taxi 2014

Schema Type: Logical Design

Description: Done Logical Design but lack of Indexes

Status: 90% Done

Last Updated: 2/1/2021

Person Responsible for Design: Nguyễn Bách Việt

Supporting team (if applicable): Nguyễn Phượng Vỹ

## DATASOURCE ANALYSIS

* 1. **Yellow Taxi**

1. **Mô tả dữ liệu:**

Dữ liệu các chuyến xe đã hoàn thành của hãng Yellow Taxi ở New York từ 2014 đến năm 2016

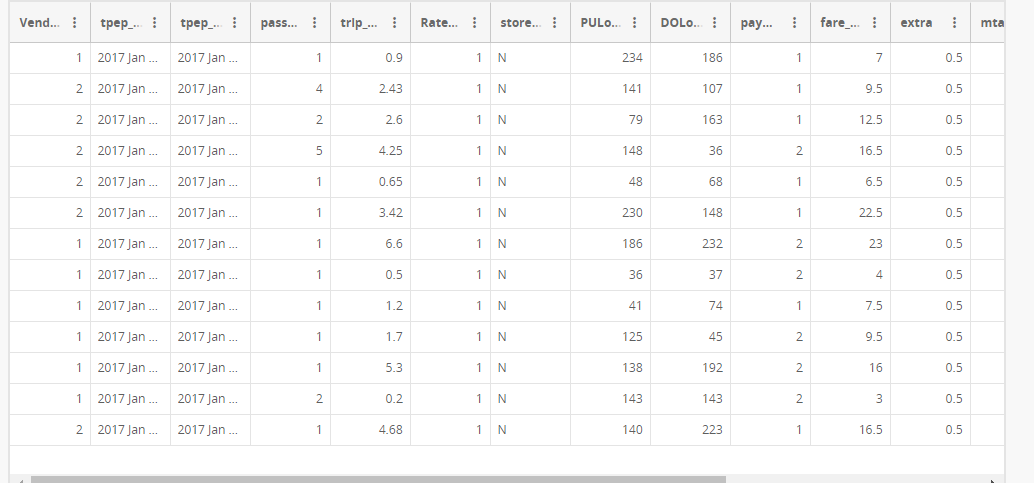
1. **Nguồn dữ liệu:**

**2016:** <https://data.cityofnewyork.us/Transportation/2016-Yellow-Taxi-Trip-Data/k67s-dv2t>

**2017:** <https://data.cityofnewyork.us/Transportation/2017-Yellow-Taxi-Trip-Data/biws-g3hs>

**2018:** https://data.cityofnewyork.us/Transportation/2018-Yellow-Taxi-Trip-Data/t29m-gskq/data

1. **Hình ảnh nguồn dữ liệu**



**Dữ liệu Yellow Taxi 2016,2017,2018**

1. **Bảng phân tích nguồn dữ liệu**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column | Key Type | Data Type | Is Nullable | Description |
| vendor\_id | none | String | Not Null | Mã các nguồn cung cấp dữ liệu yellow taxi (CMT=Creative Mobile Technologies VTS= VeriFone, Inc. DDS=Digital Dispatch Systems) |
| tpep\_pickup\_datetime | None | Datetime | Not Null | Thời gian đón khách |
| tpep\_dropoff\_datetime | None | Datetime | Not Null | Thời gian trả khách |
| passenger\_count | None | Int | Not Null | Số lượng khách trên xe (nhập tự động). |
| trip\_distance | None | Float | Not Null | Quãng đường di chuyển. |
| pulocationid | None | Double | Not Null | LocationID điểm đón khách. |
| store\_and\_fwd\_flag | None | Boolean | Not Null | Cho biết record về chuyến đi có được lưu trong bộ nhớ của xe không (Y = Có, N = Không) |
| dolacationid | None | Double | Not Null | Locationid điểm trả khách. |
| payment\_type | None | String | Not Null | Loại thanh toán khách trả cho chuyến đi. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip (1 = Thẻ tín dụng, 2 = Tiền mặt, 3 = Không tính phí, 4 = Tranh chấp, 5 = Không xác định, 6 = Chuyến đi bị hủy) |
| fare\_amount | None | Float | Not Null | Đơn giá |
| MTA\_tax | None | Float | Not Null | Thuế MTA $ 0,50 được tự động kích hoạt dựa trên tỷ giá đo được sử dụng. |
| tip\_amount | None | Float |  | Tiền tip |
| tolls\_amount | None | Float |  | Phí cầu đường |
| total\_amount | None | Float |  | Tổng tất cả chi phí khách phải trả (không bao gồm tiền tip, im\_surchagre) |
| imp\_surcharge | None | Float |  | $0.3 phụ phí cải thiện chuyến đi (bắt đầu tính từ 2015) |
| extra | None | Float |  | Phụ phí (gồm $0,5và $1 cho giờ cao điểm và phí qua đêm) |
| rate\_code | None | Int | Not Null | Mã giá cuối cùng có hiệu lực vào cuối chuyến đi (1 = Giá tiêu chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Giá thỏa thuận 6 = Đi theo nhóm) |

* 1. **New York City Census Tracts for 2010 US Census**

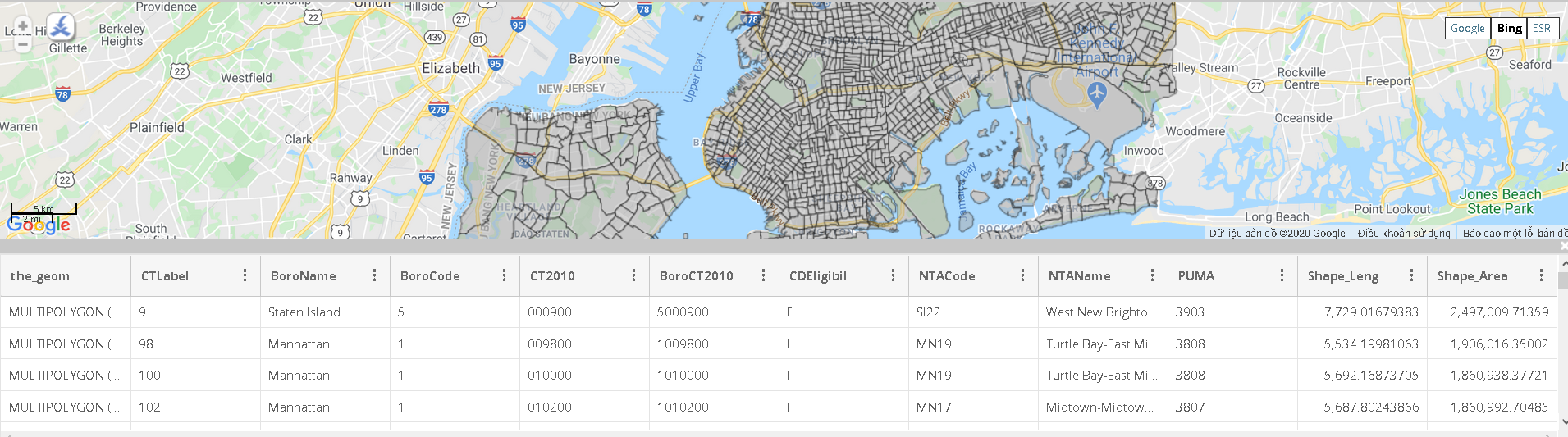
1. **Mô tả dữ liệu**

Các phân tích điều tra dân số cho cuộc điều tra dân số Hoa Kỳ năm 2010. Các tệp ranh giới này được dẫn xuất từ dự án TIGER của Cục điều tra dân số Hoa Kỳ và đã đực sửa đổi về mặt địa lý để phù hợp với bản đồ cơ sơ Thành phố New York.

1. **Nguồn dữ liệu**

[*https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku*](https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku)

1. **Hình ảnh nguồn dữ liệu**



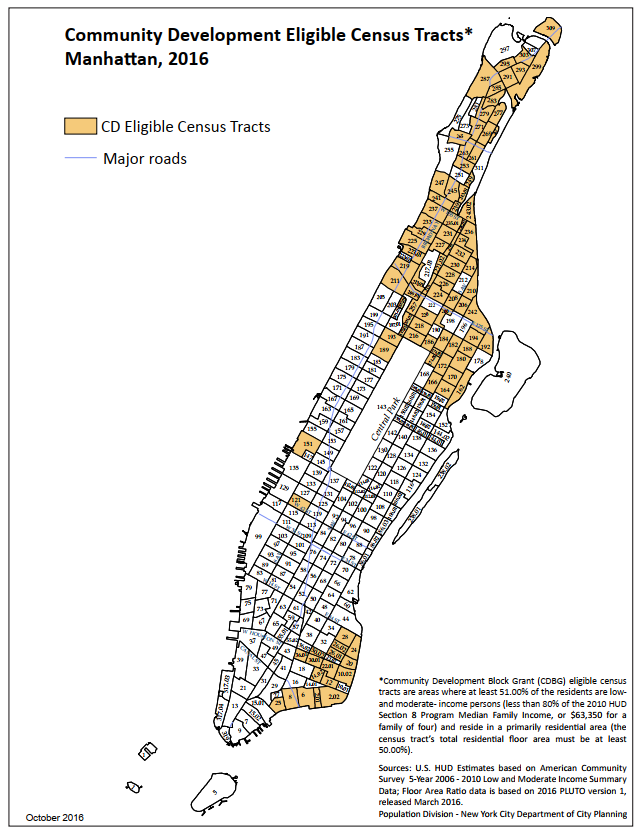
1. **Bảng phân tích nguồn dữ liệu**

Nguồn:*https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Field Name | Data type | Width | Description |
| 1 | the\_geom | string |  | Tọa độ  West: -65.257159  East: -64.699215  North: 32.915568  South: 32.495992 |
| 2 | CTlabel | string | 4 | Nhãn điều tra dân số, mỗi nhãn điều tra dân số này là duy nhất đối với mỗi Quận |
| 3 | BoroName | string | 32 | Tên Quận  List các giá trị:  Bronx, Manhattan, Queens, , Brooklyn, Staten Island |
| 4 | BoroCode | string | 1 | Tương ứng với mỗi quận là Mã quận.  Ví dụ  1 = Manhattan  2 = Bronx |
| 5 | CT2010 | string | 6 | Nhãn điều tra dân số năm 2010. Dãy số này được tạo từ “Ctlabel” + “00”  ví dụ: CTLabel: 16 , CT2010:160  CTLabel:230 ,CT2010:23000 |
| 6 | BoroCT2010 | string | 7 | Chuỗi ghép giữa BoroughCode và CT2010  Số gồm 7 chữ số trái qua phải được tạo ra từ mã quận và mã điều tra, thêm số 0 vào mã quận sao cho đủ 7 chữ số  Ví dụ:  Staten Island 5 + CTLabel 900: 5000900  Manhattan 1 + CTLabel 10200: 1010200 |
| 7 | CDEligibil | string | 1 | Community Development Block Grant Eligibility:  Tính đủ điều kiện để cấp Khối Phát triển Cộng đồng  E = Eligible: Đạt  I = Ineligible: Không Đạt |
| 8 | NTAcode | string | 4 | Neighborhood Tabulation Area Code:   * Mã khu phố, tương ứng với mỗi quận, lấy từ một hay hai kí tự đầu tiên của quận. * Ví dụ: SI07 = thuộc quận Staten Island   BK78 = thuộc quận Brooklyn  MN24 = thuộc quận Manhattan  BX08 = thuộc quận Bronx  QN51 = thuộc quận Queens |
| 9 | NTAname | string | 75 | Tên khu phố |
| 10 | PUMA | string | 4 | Public User Mircodata Area:   * Một số liệu dùng cho thống kê dân số, một PUMA là khu vực có khoảng ít nhất 100,000 người.   - Một PUMA sẽ bao gồm nhiều Community District, NTAs  - Ví dụ:  +West New Brighton-New Brighton-St. +George thuộc PUMA: 3903  +Stapleton-Rosebank thuộc PUMA: 3903 |
| 11 | Shape\_leng | double | 8 | Chiều dài khu vực |
| 12 | Shape\_area | double | 8 | Diện tích khu vực |

1. **Giải thích một số thuộc tính**

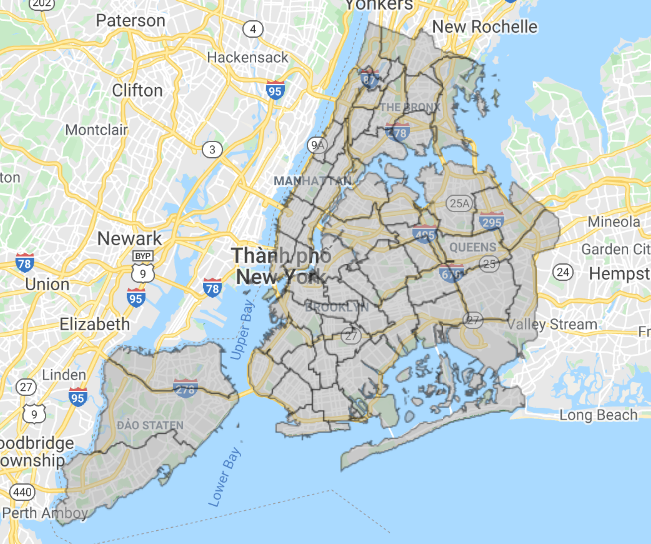
* **CDEligibil**

**Hình ảnh minh họa những khu vực đạt chuẩn là thành phố nông thôn của New York** 

Nguồn:*https://www1.nyc.gov/assets/planning/download/pdf/data-maps/maps geography/cdbg/cd\_eligibility201115\_mn2020.pdf*

* **PUMA**

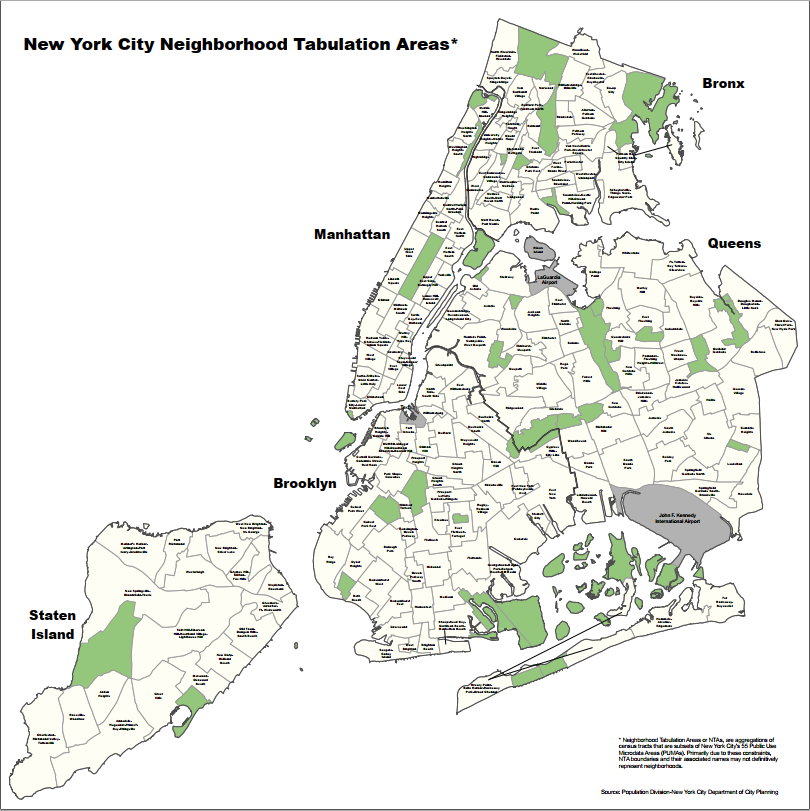
**Hình ảnh minh họa những Public user Mircodata Area**



Nguồn: *https://data.cityofnewyork.us/Housing-Development/Public-Use-Microdata-Areas-PUMA-/cwiz-gcty*

* **NTA**

**Hình ảnh minh họa những Neighborhood Tabulation Area Code của New York**

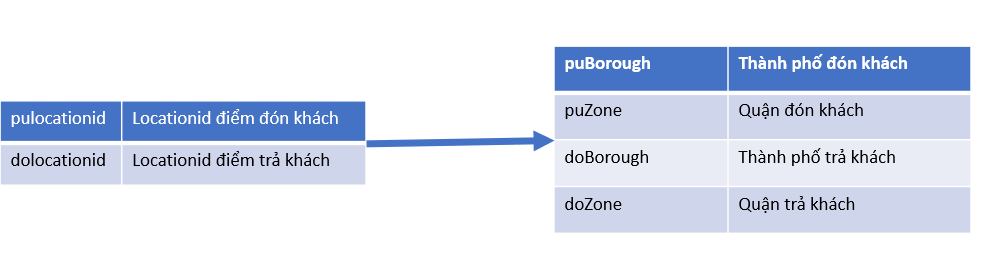


Nguồn: *https://www1.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/census2010/ntas.pdf*

### 

## YÊU CẦU CẦN SỬ DỤNG CHUYỂN ĐỔI DỮ LIỆU

1. **Từ pulocationid, dolocationid xác định thành phố, quận.**



#### **File sử dụng:**

**LocationID\_Area.csv:**

* Nguồn API:
  + *https://revgeocode.search.hereapi.com/v1/revgeocode?at=<LATITUDE>%2C<LONGITUDE>&lang=en-US&apikey=<API-KEY >*
* Các tham số truyền vào:
  + <LATITUDE>: kinh độ muốn tìm.
  + <LONGTITUDE>: vĩ độ muốn tìm.
  + <API-KEY>: Mã được cung cấp khi đăng kí dịch vụ sử dụng api.
* Hình ảnh kết quả trả về khi gọi mẫu một API với kinh độ vĩ độ:
  + Lat: 40.774415
  + Long: -73.947565

1. **Bảng phân tích dữ liệu trả về**

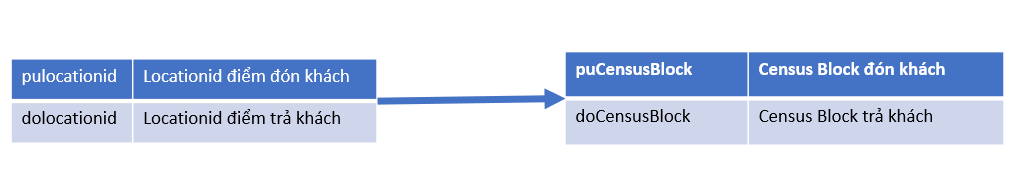
|  |  |  |  |
| --- | --- | --- | --- |
| STT | Trường | Kiểu DL | Miêu Tả |
| 1 | Locationid | String | locationid |
| 2 | Borough | String | Thành phố |
| 3 | Zone | String | Quận |
| 4 | Service\_Zone | String | Dịch vụ zone |

1. **Quá trình tổng hợp dữ liệu**

1. Lấy ra lần lượt từng cặp tọa độ pulocation/dolocation đưa vào so sánh với locationid trong file LocationID\_Area.csv để lấy được cặp pu/doBorough, pu/doZone tương ứng.

Cách thực hiện:

1. Mở <https://colab.research.google.com/notebooks/intro.ipynb#recent=true>
2. Upload file CallAPIYellowTaxi2016pynb
3. Chạy tuần tự chương trình
4. **Từ pu/dolocationid xác định Census Block ID (FIPS code)**



1. **API sử dụng:**

**FCC AreaAPI**

* Nguồn API:
  + [*https://geo.fcc.gov/api/census/block/find?latitude=<* LATITUDE *>&longitude=<* LONGTITUDE*?>&showall=true&format=json*](https://geo.fcc.gov/api/census/block/find?latitude=%3c%20LATITUDE%20%3e&longitude=%3c%20LONGTITUDE?%3e&showall=true&format=json)
* Các tham số truyền vào:
  + <LATITUDE>: kinh độ muốn tìm.
  + <LONGTITUDE>: vĩ độ muốn tìm.
* Hình ảnh kết quả trả về khi gọi mẫu một API với kinh độ vĩ độ:
  + Lat: 40.774415
  + Long: -73.947565



1. **Bảng phân tích dữ liệu trả về**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Trường cha | Trường con | Kiểu DL | Miêu Tả |
| 1 | FIPS |  | String | Mã FIPS |
| 2 | County | FIPS | String | Mã FIPS của hạt |
|  |  | Name | String | Tên hạt |
| 3 | State | FIPS | String | Mã FIPS bang |
|  |  | code | String | Mã bang |
|  |  | name | String | Tên bang |

Ta có mã phía trên: “360610144013000”

Trong đó: Mã Hạt: 36061 (5 số đầu)

Mã Bang: 36 (2 số đầu)

Census Track: 014401(6 số tiếp theo)

Census Block: 3000 (4 số cuối)

1. **Quá trình tổng hợp dữ liệu**

1. Lấy ra lần lượt từng pulocation/dolocationid đưa vào so sánh với locatoinid\_i của file nyc\_locationid\_long/lat để đổi locationid thành tọa độ.

2. Sử dụng cặp long/lat vừa thu được đường link FCC AreaAPI lấy chuỗi FIPS.

3. Cắt chuỗi FIPS từ vị trí [5:11] và bỏ các số 0 phía trước từ chuỗi FIPS vừa thu được để có Census\_BlockID

4. Thêm dữ liệu lấy được gán vào cột PUCensus\_Tract, DOCensus\_Tract

Cách thực hiện:

1. Mở <https://colab.research.google.com/notebooks/intro.ipynb#recent=true>
2. Upload file ConvertLocationIDtoCensusTract.ipynb
3. Chạy tuần tự chương trình

## KIẾN TRÚC DATA MART BÊN TRONG

* **BUSINESS REQUIREMENTS (**Danh sách các chức năng cần thiết, để thiết kế dữ liệu cho phù hợp)

|  |  |
| --- | --- |
| ID | Yêu cầu chức năng |
| REQ01 | Người dùng nghiệp vụ cần phân tích dữ liệu (số lượng đón xe (pickups) và số lượng trả khách (drop-offs)) của các chuyến đi đã được thực hiện theo thời gian (tháng, quý, năm) ở các vùng (district, city). |
| REQ02 | Người dùng nghiệp vụ cần phân tích về việc kinh doanh từ dịch vụ taxi => Cần biết doanh thu trong năm theo quận (district). |
| REQ03 | Người dùng nghiệp vụ cần phân tích thời điểm nào trong ngày là giờ cao điểm và thấp điểm của các chuyến taxi. Cũng như cần bảng báo cáo số lượng đón taxi theo giờ trong ngày trong tất cả các năm. |
| REQ04 | Người dùng nghiệp vụ cần phân tích dữ liệu (số lượng đón xe (pickups) và số lượng trả khách (drop-offs)) của các chuyến đi đã được thực hiện theo thời gian (tháng, quý, năm) ở các Census Block. |
| REQ05 | Người dùng nghiệp vụ cần phân tích số lượng khách hàng trong năm của 1 vùng (district, city) theo loại thanh toán. |
| REQ06 | Người dùng nghiệp vụ cần dự đoán cho số tiền boa (tip\_amount) theo tỉ lệ phần trăm của tổng giá vé (total\_amount) |
| REQ07 | Người dùng nghiệp vụ cần dự đoán số lượng khách theo theo 1 thời điểm cụ thể trong năm của 1 quận |

**Từ danh sách các chức năng hệ thống, ta sẽ xây dựng các mô hình Fact, Dim tương ứng phù hợp đáp ứng.**

### ***REQ01****:* Người dùng nghiệp vụ cần phân tích dữ liệu (số lượng đón xe (pickups) và số lượng trả khách (drop-offs)) của các chuyến đi đã được thực hiện theo thời gian (tháng, quý, năm) ở các vùng (district, city).

1. **Phân tích yêu cầu**
2. **Sự kiện**:

Số chuyến taxi được đặt trong tháng, quý, năm theo khu vực địa lý (district, city).

1. **Bối cảnh sự kiện**:

* Ai: Nguồn cung cấp dữ liệu (số lượng đón xe (pickups) và số lượng trả khách (drop-offs)).
* Ở đâu: Các địa điểm (district, city) trong thành phố New York.
* Cái gì: Số dịch vụ cuốc xe taxi được đặt bởi khách hàng.
* Khi nào: Các thời điểm trong tháng, quý, năm

1. **Đo lường (dữ kiện):**

Số lượng, thời gian, vị trí địa lý

1. **Phân tích yêu cầu về dữ liệu**

* Các giá trị có sẵn từ nguồn:
* pickup\_datetime
* dropoff\_datetime
* pulocationid
* dolocationid
* passenger\_count
* Các giá trị phải tính toán:
* PUBorough, PUZonedựa vào locationid
* DOBorogh, DODZone dựa vào locationid
* Cấp chi tiết dữ liệu (độ mịn):
* Một dòng trong fact tương ứng với một cuốc xe taxi được đặt bởi khách hàng vào thời gian và địa điểm đón trả cụ thể.
* **COLUMNS**

Table 1: Table (or Views) & Columns

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table | Column | Key Type | Data Type | Is Nullable | Description |
| Time | Month | None | Int | Not Null | Tháng |
|  | FiscalQuarter | None | Int | Not Null | Quý (4 tháng) |
|  | FiscalSemester | None | Int | Not Null | Kỳ (6 tháng) |
|  | FiscalYear | None | Int | Not Null | Năm (12 tháng) |
| Location | Location\_key | PrimaryKey | String | Not Null | Khóa chính |
|  | District | None | String | Not Null | Tên quận |
|  | City | None | String | Not Null | Tên thành phố |
| NguonDuLieu | ID | PrimaryKey | Int | Not Null | Mã nguồn dữ liệu |
|  | NguonDuLieu | None | String |  | Tên nguồn dữ liệu |
| TrangThai | StatusID | PrimaryKey | Int | Not Null | Mã trạng thái |
|  | Status | None | String |  | Tên trạng thái |
| TripRecords | Triprecord\_key | ForeignKey | String | Not Null | Khóa chính |
|  | TimeKey\_PU | ForeignKey | String | Not Null | Thời gian đón khách |
|  | LocationKey\_PU | ForeignKey | String | Not Null | Vị trí đón khách. |
|  | TimeKey\_DO | ForeignKey | String | Not Null | Thời gian trả khách |
|  | LocationKey\_DO | ForeignKey | String | Not Null | Vị trí trả khách. |
|  | passenger\_count | None | Int | Not Null | Số lượng khách trên xe. |
|  | NguonDuLieu | ForeignKey | Int | Not Null | Nguồn dữ liệu |
|  | Status | ForeignKey | Int | Not Null | Trạng thái |

* **CONSTRAINTS**

Table 2: Constraints

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Constraint  Name | Constraint  Type | Table | Column | Description |
| TimeKey\_FK | PrimaryKey | Time | Month, FiscalQuarter, FiscalSemester, FiscalYear | Khóa ngoại bảng Time |
| LocationKey\_FK | PrimaryKey | Location | District, Street | Khóa ngoại bảng Location |
| NguonDuLieu\_FK | PrimaryKey | NguonDuLieu | NguonDuLieu | Khóa ngoại bảng NguonDuLieu |
| TrangThai\_FK | PrimaryKey | TrangThai | Status | Khóa ngoại bảng Status |

#### 

#### **Xác định dimension table + SCD:**

Các chiều liên quan đến sự kiện phân tích :

* Time dimesion :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Location dimension :

+ Có thể thay đổi, các thành phố, quận hay khu vực có thể hợp nhất lại hoặc tách ra (hiếm xảy ra).

+ SCD loại 3 -> không thay đổi nhiều thuộc tính

* TripRecords dimension :

Các thuộc tính thường xuyên thay đổi SCD :

+ passenger\_count

* NguonDuLieu dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Status dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

**DATA STORE LOGICAL MODEL**

// UPDATE

### ***REQ02****:* Người dùng nghiệp vụ cần phân tích về việc kinh doanh từ dịch vụ taxi => Cần biết doanh thu trong năm theo quận (district).

1. **Phân tích yêu cầu**
2. **Sự kiện:**

1 khách hàng đặt 1 cuốc xe taxi

1. **Bối cảnh sự kiện:**  
   Ai: Nguồn cung cấp dữ liệu (doanh thu = Total\_amount )

Ở đâu: Các địa điểm (quận) trong thành phố New York

Cái gì: Dịch vụ cuốc xe taxi

Khi nào: Ngày đặt dịch vụ

1. **Đo lường (dữ kiện):**

Doanh thu, thời gian (năm), vị trí địa lý (quận)

* Các giá trị có sẵn từ nguồn:
* pickup\_datetime
* dropoff\_datetime
* pulocationid
* dolocationid
* passenger\_count
* Fare\_amount
* Extra
* MTA\_tax
* Imp\_surcharge
* Tip\_amount
* Tolls\_amount
* Total\_amount

1. **Phân tích yêu cầu về dữ liệu.**

* Các giá trị phải tính toán:
* Total\_amount

Total\_amount = Extra + fare\_amount + mta\_tax + imp\_surcharge + tolls\_amount

* Cấp chi tiết dữ liệu (độ mịn):

Một dòng trong fact tương ứng mỗi dịch vụ cuốc xe được đặt và thanh toán bởi khách hàng.

* Nhu cầu phức tạp hơn:

Sự thay đổi theo thời gian của các loại chi phí như chi phí cơ sở hạ tầng; phí dịch vụ giờ cao điểm, qua đêm…

* COLUMNS:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table | Column | Key Type | Data Type | Is Nullable | Description |
| TimeKey | FiscalYear | None | Int | Not Null | Năm (12 tháng) |
| Location | Location\_key | PrimaryKey | String | Not Null | Khóa chính |
|  | District | None | String | Not Null | Tên quận |
|  | Street | None | String | Not Null | Tên đường |
| NguonDuLieu | ID | PrimaryKey | Int | Not Null | Mã nguồn dữ liệu |
|  | NguonDuLieu | None | String |  | Tên nguồn dữ liệu |
| TrangThai | StatusID | PrimaryKey | Int | Not Null | Mã trạng thái |
|  | Status | None | String |  | Tên trạng thái |
| TripRecords | Triprecord\_key | ForeignKey | String | Not Null | Khóa chính |
|  | TimeKey\_PU | ForeignKey | String | Not Null | Thời gian đón khách |
|  | LocationKey\_PU | ForeignKey | String | Not Null | Vị trí đón khách. |
|  | TimeKey\_DO | ForeignKey | String | Not Null | Thời gian trả khách |
|  | LocationKey\_DO | ForeignKey | String | Not Null | Vị trí trả khách. |
|  | total\_amount | None | Float | Not Null | Tổng tất cả chi phí khách phỉa trả (không bao gồm tiền tip, im\_surchagre) |
|  | NguonDuLieu | ForeignKey | Int | Not Null | Nguồn dữ liệu |
|  | Status | ForeignKey | Int | Not Null | Trạng thái |

* CONTRAINTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Constraint  Name | Constraint  Type | Table | Column | Description |
| TimeKey\_FK | PrimaryKey | Time | Month, FiscalQuarter, FiscalSemester, FiscalYear | Khóa ngoại bảng Time |
| LocationKey\_FK | PrimaryKey | Location | District, Street | Khóa ngoại bảng Location |
| NguonDuLieu\_FK | PrimaryKey | NguonDuLieu | NguonDuLieu | Khóa ngoại bảng NguonDuLieu |
| TrangThai\_FK | PrimaryKey | TrangThai | Status | Khóa ngoại bảng Status |

1. **Xác định Dimension table + SCD:**

Các chiều liên quan đến sự kiện phân tích :

* Time dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Location dimension :

+ Tương lai có thể thay đổi, các thành phố, quận hay khu vực có thể hợp nhất lại hoặc tách ra (hiếm xảy ra).

+ SCD loại 3 -> không thay đổi nhiều thuộc tính

* TripRecords dimension :

Các thuộc tính thường xuyên thay đổi RCD :

+ Total\_amount

* NguonDuLieu dimension :

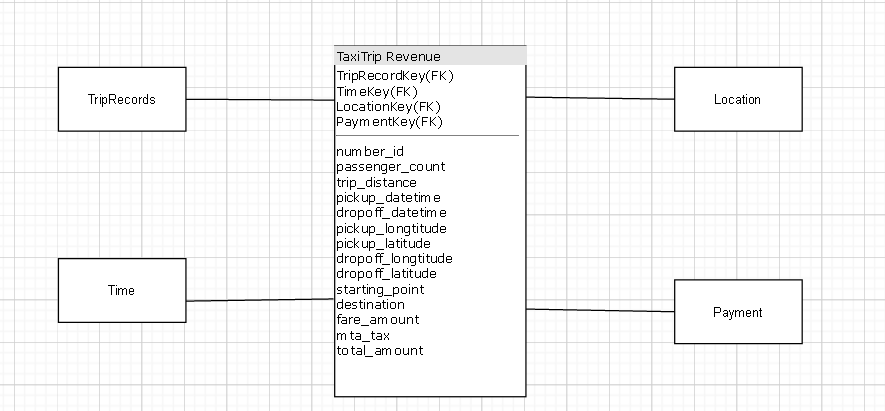
+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Status dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

**DATA STORE LOGICAL MODEL**

// UPDATE



### ***REQ03****:* Người dùng nghiệp vụ cần phân tích thời điểm nào trong ngày là giờ cao điểm và thấp điểm của các chuyến taxi. Cũng như cần bảng báo cáo số lượng đón taxi theo giờ trong ngày trong tất cả các năm.

1. **Phân tích yêu cầu**
2. **Sự kiện**:

Số lượng taxi được đón theo giờ trong ngày của các năm, xác định giờ cao điểm và thấp điểm.

1. **Bối cảnh sự kiện**:

* Ai: Nguồn cung cấp dữ liệu
* Ở đâu: Các địa điểm trong thành phố New York
* Cái gì: Số lượng xe taxi đặt nhiều nhất và thấp nhất trong 1 thời điểm
* Khi nào: Mỗi giờ trong ngày

1. **Đo lường (dữ kiện):**

Số lượng, thời gian

1. **Phân tích yêu cầu về dữ liệu**

* Các giá trị có sẵn từ nguồn:
* pickup\_datetime
* pulocationid
* Các giá trị phải tính toán:
* Số lượng đón taxi trong 1 giờ (TotalTaxi)
* Cấp chi tiết dữ liệu (độ mịn):
* Một dòng trong fact tương ứng với một cuốc xe taxi được đặt bởi khách hàng vào thời gian cụ thể.
* **COLUMNS**

Table 1: Table (or Views) & Columns

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table | Column | Key Type | Data Type | Is Nullable | Description |
| TimeKey | Hour | None | Datetime | Not Null | Dữ liệu giờ |
|  | Day | None | Datetime | Not Null | Dữ liệu ngày |
|  | Week | None | Int | Not Null | Tuần thứ mấy của tháng |
|  | Month | None | Int | Not Null | Tháng |
|  | FiscalYear | None | Int | Not Null | Năm (12 tháng) |
| Location | Location\_key | PrimaryKey | String | Not Null | Khóa chính |
|  | District | None | String | Not Null | Tên quận |
|  | Street | None | String | Not Null | Tên đường |
| NguonDuLieu | ID | PrimaryKey | Int | Not Null | Mã nguồn dữ liệu |
|  | NguonDuLieu | None | String |  | Tên nguồn dữ liệu |
| TrangThai | StatusID | PrimaryKey | Int | Not Null | Mã trạng thái |
|  | Status | None | String |  | Tên trạng thái |
| TripRecords | Triprecord\_key | ForeignKey | String | Not Null | Khóa chính |
|  | TimeKey\_PU | ForeignKey | String | Not Null | Thời gian đón khách |
|  | LocationKey\_PU | ForeignKey | String | Not Null | Vị trí đón khách. |
|  | TotalTaxi | None | int | NotNull | Số lượng đón taxi trong 1 giờ |
|  | NguonDuLieu | ForeignKey | Int | Not Null | Nguồn dữ liệu |
|  | Status | ForeignKey | Int | Not Null | Trạng thái |

* **CONSTRAINTS**

Table 2: Constraints

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Constraint  Name | Constraint  Type | Table | Column | Description |
| TimeKey\_FK | PrimaryKey | Time | Hour | Khóa ngoại bảng Time |
| LocationKey\_FK | PrimaryKey | Location | District, Street | Khóa ngoại bảng Location |
| NguonDuLieu\_FK | PrimaryKey | NguonDuLieu | NguonDuLieu | Khóa ngoại bảng NguonDuLieu |
| TrangThai\_FK | PrimaryKey | TrangThai | Status | Khóa ngoại bảng Status |

#### **Xác định dimension table + SCD:**

Các chiều liên quan đến sự kiện phân tích :

* Time dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Location dimension :

+ Tương lai có thể thay đổi, các thành phố, quận hay khu vực có thể hợp nhất lại hoặc tách ra (hiếm xảy ra).

+ SCD loại 3 -> không thay đổi nhiều thuộc tính

* TripRecords dimension :

Các thuộc tính thường xuyên thay đổi RCD :

+ TotalTaxi

* NguonDuLieu dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Status dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

### ***REQ04****:* Người dùng nghiệp vụ cần phân tích dữ liệu (số lượng đón xe (pickups) và số lượng trả khách (drop-offs)) của các chuyến đi đã được thực hiện theo thời gian (tháng, quý, năm) ở các Census Block. (chưa ổn)

1. **Phân tích yêu cầu**
2. **Sự kiện**:

Số dịch vụ chuyến xe taxi được đặt vào thời gian cụ thể trong tháng, quý, năm theo các Census Block.

1. **Bối cảnh sự kiện**:

* Ai: Nguồn cung cấp dữ liệu (số lượng đón xe (pickups) và số lượng trả khách (drop-offs)).
* Ở đâu: Các địa điểm (theo Census Block) trong thành phố New York.
* Cái gì: Số dịch vụ cuốc xe taxi được đặt bởi khách hàng.
* Khi nào: Các thời điểm trong tháng, quý, năm,

1. **Đo lường (dữ kiện):**

Số lượng, thời gian, vị trí địa lý

1. **Phân tích yêu cầu về dữ liệu**

* Các giá trị có sẵn từ nguồn:
* pickup\_datetime
* dropoff\_datetime
* pulocationid
* dolocationid
* passenger\_count
* passenger\_count
* Các giá trị phải tính toán:
* Starting Point(điểm xuất phát) dựa vào tọa độ
* Destination(nơi đến) dựa vào tọa độ
* Cấp chi tiết dữ liệu (độ mịn):
* Một dòng trong fact tương ứng với một cuốc xe taxi được đặt bởi khách hàng vào thời gian và địa điểm đón trả cụ thể.
* **COLUMNS**

Table 1: Table (or Views) & Columns

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table | Column | Key Type | Data Type | Is Nullable | Description |
| Time | Month | None | Int | Not Null | Tháng |
|  | FiscalQuarter | None | Int | Not Null | Quý (4 tháng) |
|  | FiscalSemester | None | Int | Not Null | Kỳ (6 tháng) |
|  | FiscalYear | None | Int | Not Null | Năm (12 tháng) |
| Census\_Track | CensusBlock\_key | PrimaryKey | String | Not Null | Khóa chính |
|  | District | None | String | Not Null | Tên quận |
|  | Street | None | String | Not Null | Tên đường |
| NguonDuLieu | ID | PrimaryKey | Int | Not Null | Mã nguồn dữ liệu |
|  | NguonDuLieu | None | String |  | Tên nguồn dữ liệu |
| TrangThai | StatusID | PrimaryKey | Int | Not Null | Mã trạng thái |
|  | Status | None | String |  | Tên trạng thái |
| TripRecords | Triprecord\_key | ForeignKey | String | Not Null | Khóa chính |
|  | TimeKey\_PU | ForeignKey | String | Not Null | Thời gian đón khách |
|  | LocationKey\_PU | ForeignKey | String | Not Null | Vị trí đón khách. |
|  | TimeKey\_DO | ForeignKey | String | Not Null | Thời gian trả khách |
|  | LocationKey\_DO | ForeignKey | String | Not Null | Vị trí trả khách. |
|  | passenger\_count | None | Int | Not Null | Số lượng khách trên xe. |
|  | NguonDuLieu | ForeignKey | Int | Not Null | Nguồn dữ liệu |
|  | Status | ForeignKey | Int | Not Null | Trạng thái |

* **CONSTRAINTS**

Table 2: Constraints

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Constraint  Name | Constraint  Type | Table | Column | Description |
| TimeKey\_FK | PrimaryKey | Time | Month, FiscalQuarter, FiscalSemester, FiscalYear | Khóa ngoại bảng Time |
| LocationKey\_FK | PrimaryKey | Location | District, Street | Khóa ngoại bảng Location |
| NguonDuLieu\_FK | PrimaryKey | NguonDuLieu | NguonDuLieu | Khóa ngoại bảng NguonDuLieu |
| TrangThai\_FK | PrimaryKey | TrangThai | Status | Khóa ngoại bảng Status |

#### **Xác định dimension table + SCD:**

Các chiều liên quan đến sự kiện phân tích :

* Time dimesion :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Location dimension :

+ Có thể thay đổi, các thành phố, hạt, quận hay khu vực có thể hợp nhất lại hoặc tách ra (hiếm xảy ra).

+ SCD loại 3 -> không thay đổi nhiều thuộc tính

* TripRecords dimension :

Các thuộc tính thường xuyên thay đổi SCD :

+ passenger\_count

* NguonDuLieu dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Status dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

**DATA STORE LOGICAL MODEL**

### ***REQ05****:* Người dùng nghiệp vụ cần phân tích số lượng khách hàng trong năm của 1 vùng (district, city) theo loại thanh toán.

1. **Phân tích yêu cầu**
2. **Sự kiện**:

Số dịch vụ chuyến xe taxi được đặt vào thời gian cụ thể trong tháng, quý, năm theo khu vực địa lý (district, city).

1. **Bối cảnh sự kiện**:

* Ai: Nguồn cung cấp dữ liệu (số lượng đón xe (pickups) và số lượng trả khách (drop-offs)).
* Ở đâu: Các địa điểm (district, city) trong thành phố New York.
* Cái gì: Số dịch vụ cuốc xe taxi được đặt bởi khách hàng.
* Khi nào: Các thời điểm trong năm.

1. **Đo lường (dữ kiện):**

Số lượng, thời gian, vị trí địa lý

1. **Phân tích yêu cầu về dữ liệu**

* Các giá trị có sẵn từ nguồn:
* pickup\_datetime
* dropoff\_datetime
* pulocationid
* dolocationid
* passenger\_count
* Payment\_type
* Các giá trị phải tính toán:
* Starting Point(điểm xuất phát) dựa vào tọa độ
* Destination(nơi đến) dựa vào tọa độ
* Cấp chi tiết dữ liệu (độ mịn):
* Một dòng trong fact tương ứng với một cuốc xe taxi được đặt bởi khách hàng vào thời gian và địa điểm đón trả cụ thể.
* **COLUMNS**

Table 1: Table (or Views) & Columns

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table | Column | Key Type | Data Type | Is Nullable | Description |
| Time | FiscalYear | None | Int | Not Null | Năm (12 tháng) |
| Location | Location\_key | PrimaryKey | String | Not Null | Khóa chính |
|  | District | None | String | Not Null | Tên quận |
|  | Street | None | String | Not Null | Tên đường |
| NguonDuLieu | ID | PrimaryKey | Int | Not Null | Mã nguồn dữ liệu |
|  | NguonDuLieu | None | String |  | Tên nguồn dữ liệu |
| TrangThai | StatusID | PrimaryKey | Int | Not Null | Mã trạng thái |
|  | Status | None | String |  | Tên trạng thái |
| Payment | PaymentID | PrimaryKey | Int | Not Null | Mã loại thanh toán |
|  | Payment\_type | None | String |  | Tên loại thanh toán |
| TripRecords | Triprecord\_key | ForeignKey | String | Not Null | Khóa chính |
|  | TimeKey\_PU | ForeignKey | String | Not Null | Thời gian đón khách |
|  | LocationKey\_PU | ForeignKey | String | Not Null | Vị trí đón khách. |
|  | TimeKey\_DO | ForeignKey | String | Not Null | Thời gian trả khách |
|  | LocationKey\_DO | ForeignKey | String | Not Null | Vị trí trả khách. |
|  | PaymentKey | ForeignKey | String | Not Null | Mã loại thanh toán. |
|  | passenger\_count | None | Int | Not Null | Số lượng khách trên xe. |
|  | NguonDuLieu | ForeignKey | Int | Not Null | Nguồn dữ liệu |
|  | Status | ForeignKey | Int | Not Null | Trạng thái |

* **CONSTRAINTS**

Table 2: Constraints

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Constraint  Name | Constraint  Type | Table | Column | Description |
| TimeKey\_FK | PrimaryKey | Time | FiscalYear | Khóa ngoại bảng Time |
| LocationKey\_FK | PrimaryKey | Location | District, Street | Khóa ngoại bảng Location |
| NguonDuLieu\_FK | PrimaryKey | NguonDuLieu | NguonDuLieu | Khóa ngoại bảng NguonDuLieu |
| TrangThai\_FK | PrimaryKey | TrangThai | Status | Khóa ngoại bảng Status |
| PaymentKey\_FK | PrimaryKey | Payment | Payment\_type | Khóa ngoại bảng Payment |

#### **Xác định dimension table + SCD:**

Các chiều liên quan đến sự kiện phân tích :

* Time dimesion :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Location dimension :

+ Có thể thay đổi, các thành phố, hạt, quận hay khu vực có thể hợp nhất lại hoặc tách ra (hiếm xảy ra).

+ SCD loại 3 -> không thay đổi nhiều thuộc tính

* TripRecords dimension :

Các thuộc tính thường xuyên thay đổi SCD :

+ passenger\_count

* NguonDuLieu dimension :

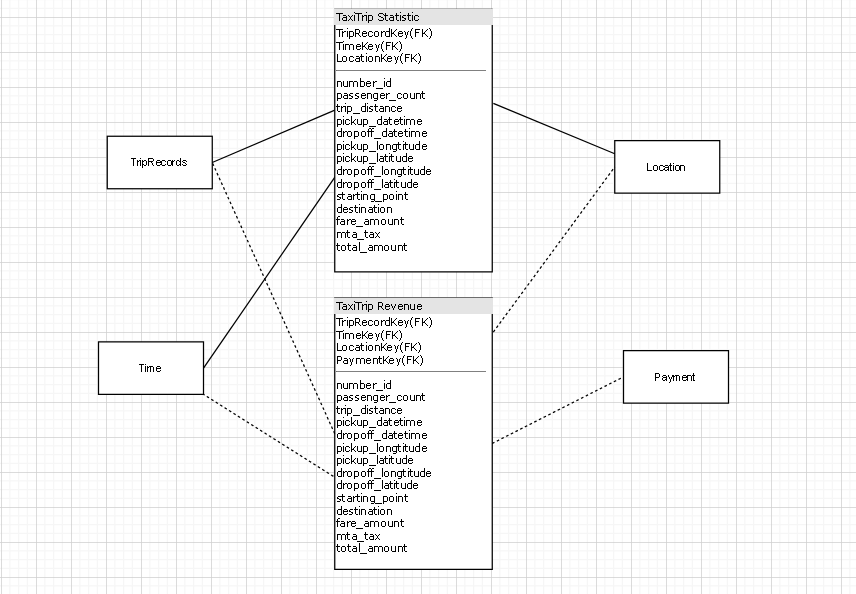
+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

* Status dimension :

+ Các giá trị không thay đổi, do người dùng nghiệp vụ sử dụng tùy chọn

**DATA STORE LOGICAL MODEL**

**MÔ HÌNH HÓA**



# Business Process

## Quy trình đặt xe Taxi

* Mô tả yêu cầu: Quy trình khách hàng đặt xe taxi.
* Nguồn dữ liệu:

Payment\_key: PaymentTypeID, PaymentTypeName

Location\_key: Street\_key, City, District, Borough

Vendor\_key: VendorID, VendorName

* Business processes – Quy trình nghiệp vụ

B1: Gọi điện hoặc đăng ký thông tin vị trí của mình theo hãng xe của Taxi

B2: Chọn loại xe cần đặt, hình thức và thông tin liên lạc

B3: Chọn phương thức thanh toán.

B4: Đặt xe.

* List business groups involved and describe type of involvement – Danh sách nhóm nghiệp vụ liên quan và mô tả loại hình tham gia
* Khách hàng: người cung cấp thông tin
* Nhân viên nhập liệu: người ghi nhận thông tin

## Quy trình điều động xe Taxi

* Mô tả yêu cầu: Quy trình điều động xe Taxi đến nơi khách cần
* Nguồn dữ liệu

Location\_key: Street\_key, City, District, Borough

Vendor\_key: VendorID, VendorName

Status\_key: StatusID, StatusName

* Business processes – Quy trình nghiệp vụ

B1: Các xe taxi tập trung tại bãi đậu xe.

B2: Nhân viên kiểm tra xe có đủ tiêu chuẩn, đáp ứng nhu cầu khách hay không

B3: Tài xế nhận thông tin đặt xe từ khách. Xác định rõ loại xe bao nhiêu chỗ, địa điểm, thời gian đón khách.

B4: Tài xế đến địa điểm và đưa đón khách theo yêu cầu.

# Thiết kế data flow

1. **Hiện trạng hệ thống**
2. Yellow Taxi là một hệ thống taxi công nghệ hoạt động tại New York.
3. Khách hàng có thể gọi xe bất cứ đâu ở New York vì lượng Yellow taxi lưu thông rất nhiều trên đường, có thể đi một mình hoặc một nhóm (nhóm thì phải <= 4 người).
4. Giá cả liên quan trực tiếp đến các mô hình nổi tiếng được quan sát trong quá khứ về sự di chuyển của con người trong thành thị.
5. Hầu hết là dữ liệu về các chuyển động của taxi nằm trong phạm vi khoảng cách ngắn.
6. Hơn nữa, phân phối tổng thể của các chuyển động trong không gian khớp trực tiếp với phân phối thống kê của giá cước taxi mà khách hàng trả.
7. Có nhiều phương thức chi trả: tiền mặt, thẻ, visa...
8. **Xác định nhu cầu**

*Một số quy tình nghiệp vụ sẵn có (Ví dụ: đặt xe, giao trả khách.)*

1. Người dùng nghiệp vụ cần phân tích thông tin đặt xe, đón trả khách. Cần biết được doanh số, chi phí, lợi nhuận theo một loại tiền tệ ($).
2. Cấp quản lý district có thể xem dữ liệu hằng ngày trong một vài tuần trước đó ở mức tổng quát hoặc chi tiết về doanh số, chi phí, lợi nhuận để hiểu nguyên nhân của việc doanh số thấp, lợi nhuận thấp, xe nào, khách hàng nào... là nguyên nhân của vấn đề.
3. **Yêu cầu về chức năng**
4. Sử dụng HERE API để lấy địa chỉ (street, district, city, state) từ tọa độ đón và trả khách.
5. Sử dụng dịch vụ hoặc thư viện để lấy Census Block ID từ tọa độ đón trả khách.
6. Phân tích chuyến đi:

* Theo Geography (street, district, state, boroughs),...
* Theo Census Block, theo thời gian.
* Theo loại thanh toán.
* Theo số tổng thanh toán (bao gồm phí giờ cao điểm, phí qua đêm, phụ phí cải thiện chuyến đi, boa,...).

1. Phân tích phân bố lượng đón khách (pickups) và số lượng trả khách (drop-offs) ở các vùng (district, city) theo thời gian tháng, quý, năm.
2. Phân tích thời điểm nào trong ngày là giờ cao điểm và thấp điểm của các chuyến taxi.
3. Vẽ đồ thị phân bố số lượng đón taxi theo giờ trong ngày trong tất cả các năm.
4. Thống kê doanh thu năm theo quận (district) đón khách.
5. Thống kê chi tiết những nơi đón trả khách nhiều nhất (thông qua kinh độ và vĩ độ) theo giờ, ngày, tháng, năm để bố trí gần đó những cửa hàng bán đồ ăn nhanh hay vật dụng cần thiết như ô khi trời mưa, nước khoáng,...
6. [Advanced] Phân tích những nơi có doanh thu, lợi nhuận nhiều nhất (thông qua kinh độ vĩ độ) theo giờ, ngày, tháng, năm để bố trí nhiều xe hơn để đáp ứng nhu cầu đi lại của khách hàng.
7. [Advanced] Phân tích những yếu tố ảnh hưởng đến trải nghiệm người dùng (giá cả, thời gian đón, chất lượng xe, sự thân thiện, tận tình,... của tài xế).
8. **Yêu cầu về dữ liệu**
9. Cấp quản lý district có thể hiểu được xu hướng chung tổng thể thông qua thông tin lộ trình đón trả khách của tất cả taxi ở district mình.
10. Hệ thống có thể hiển thị các hình, biểu đồ cho phép in được, có thể xuất ra excel hoặc csv, pdf,...
11. **Yêu cầu về hệ thống**
12. **Yêu cầu về giao diện**

* Giao diện thiết kế dễ nhìn, dễ sử dụng, bố cục rõ ràng.
* Thông tin các tuyến đường hiện lên rõ ràng và chi tiết, có hình ảnh taxi + tài xế kèm theo.
* Nâng cấp giao diện mới.
* Hình ảnh, video up lên có độ chất lượng cao.

1. **Yêu cầu về xử lý:**

Hệ thống được xử lý với bất kì trình diện Web nào.

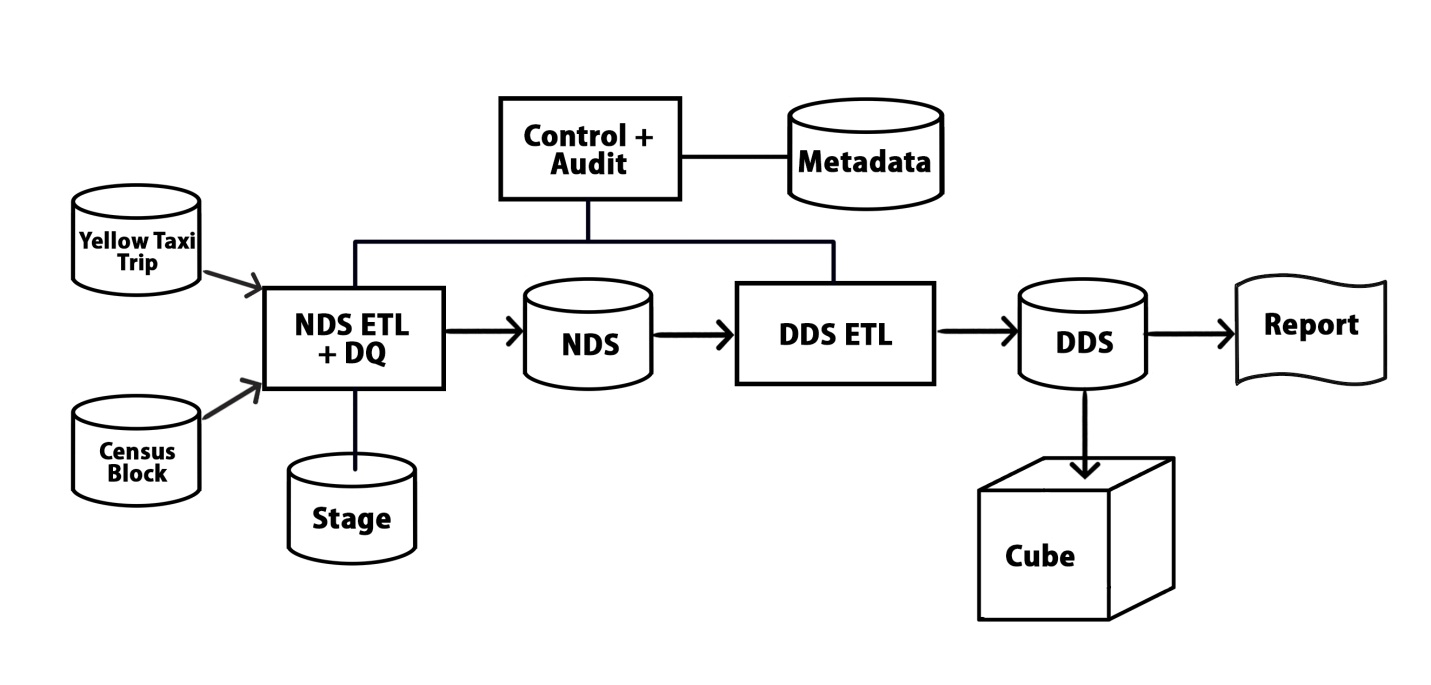
1. **Yêu cầu về hiệu suất**

* Thao tác xử lý nhanh không quá 15s
* Hệ thống nhận thông tin, kiểm kê cập nhật ngày
* Cần xử lý khi hệ thống quá tải người dùng.
* Tối ưu hóa mặt kĩ thuật.
* Các thao tác xử lý đơn giản.

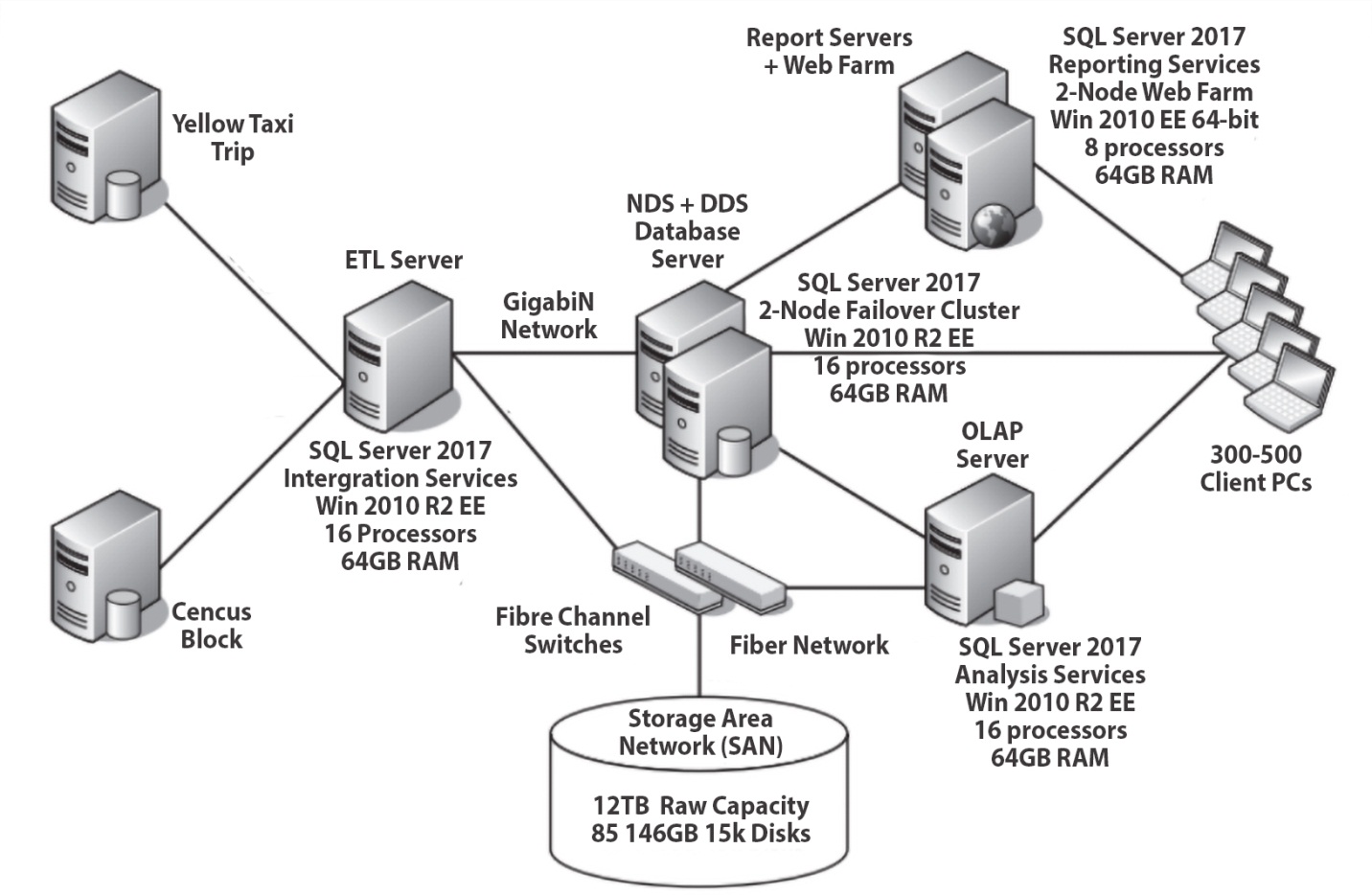
1. **Yêu cầu về bảo mật**

* Dữ liệu giao dịch, cá nhân được bảo mật.
* Dữ liệu các dự án nguồn chỉ có cấp Giám đốc mới xem được
* Chỉ có quản lý mới xem được thống kê doanh thu

1. **Có kèm bảng hướng dẫn sử dụng**
2. **Kiến trúc luồng dữ liệu**



# Thiết kế KIẾN TRÚC DỮ LIỆU



# Approvals

|  |  |  |
| --- | --- | --- |
| Title | Name | Approval  Date |
| 1. Phân tích Datasource | Việt | 20/11/2020 |
| 2. Lập Fact Table | Việt | 20/11/2020 |
| 3. Lập các Dimesion Table | Việt, Vũ | 20/11/2020 |
| 4. Kiểm tra lại các Table | Việt, Vũ, Vỹ | 20/11/2020 |
| 5. Chỉnh sửa lại Fact\_TaxiReport Table | Việt, Vỹ | 20/11/2020 |
| 6. Thêm PickUpTimeKey, DropOffTimeKey từ TimeKey Dimension Table | Việt, Vũ, Vỹ | 20/11/2020 |
| 7. Thêm LocationKey và TimeKey Dimension Table | Việt, Vũ, Vỹ | 20/11/2020 |

# Vấn đáp - zoom

<https://drive.google.com/drive/folders/1XtulIXsCcAUiM2szG8LI7Op748fgKprG?fbclid=IwAR00dHr90Fbuvyu2R5N8DlXhygXmHe5peQSnsNp-AGvZVmqDwjV2NjEEwVg>