Building NYC Model

The main objective of this data project is to provide a data platform or perhaps a data model the most important part of developing software and ensuring a way on how NYC leadership team can think about solving the problem.

Prominent Data Points
Scope of the Work
Design Principles
Conceptual Data Model
Logical Data Model
Tables
ETL Processes
Reporting Layer
Sample PowerBI Visuals
SQL Scripting
GitHub Repository
References

Prominent Data Points

There are two important features of data that will be interesting for NYC team to look out for

- Geographical Features
  - Where ? Opportunity to Pickup Most People
  - Which ? Area pays off with Profitable Fare
  - What ? Profitable Source & Destination Pairs
- Time Based Features
  - When? Interval Peak trips are happening
  - When? Possible time giving Profitable Fare
  - What? Best time for pickup's & drops

Scope of the Work

This work covered the model to build solution to answer the above data points and not attempted to perform any data analytics part as we have plenty of work covered on internet for those things.

I am exploring few design decisions to cover for the scenarios.

Design Principles

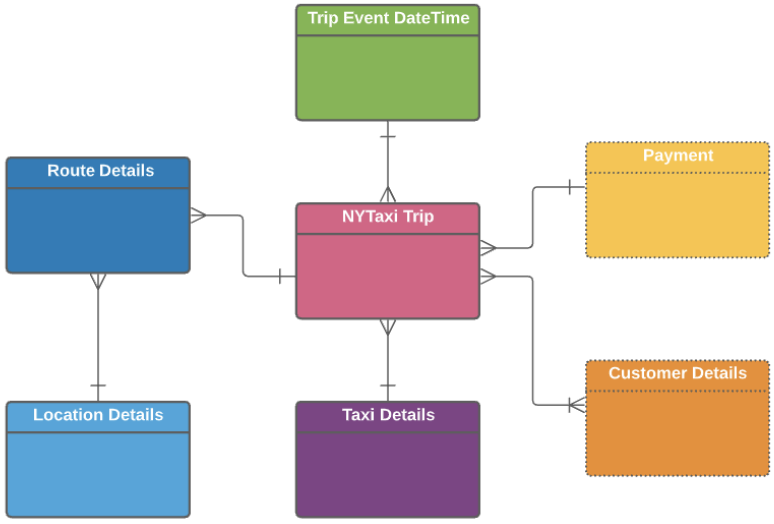
▼ Set a design goal to make the application or data project easier to change by making each table expandable by itself.

- Click the arrow again to hide this content.
- Create a toggle by typing `/toggle` and pressing `enter`.
- You can add anything to toggles, including images and embeds.

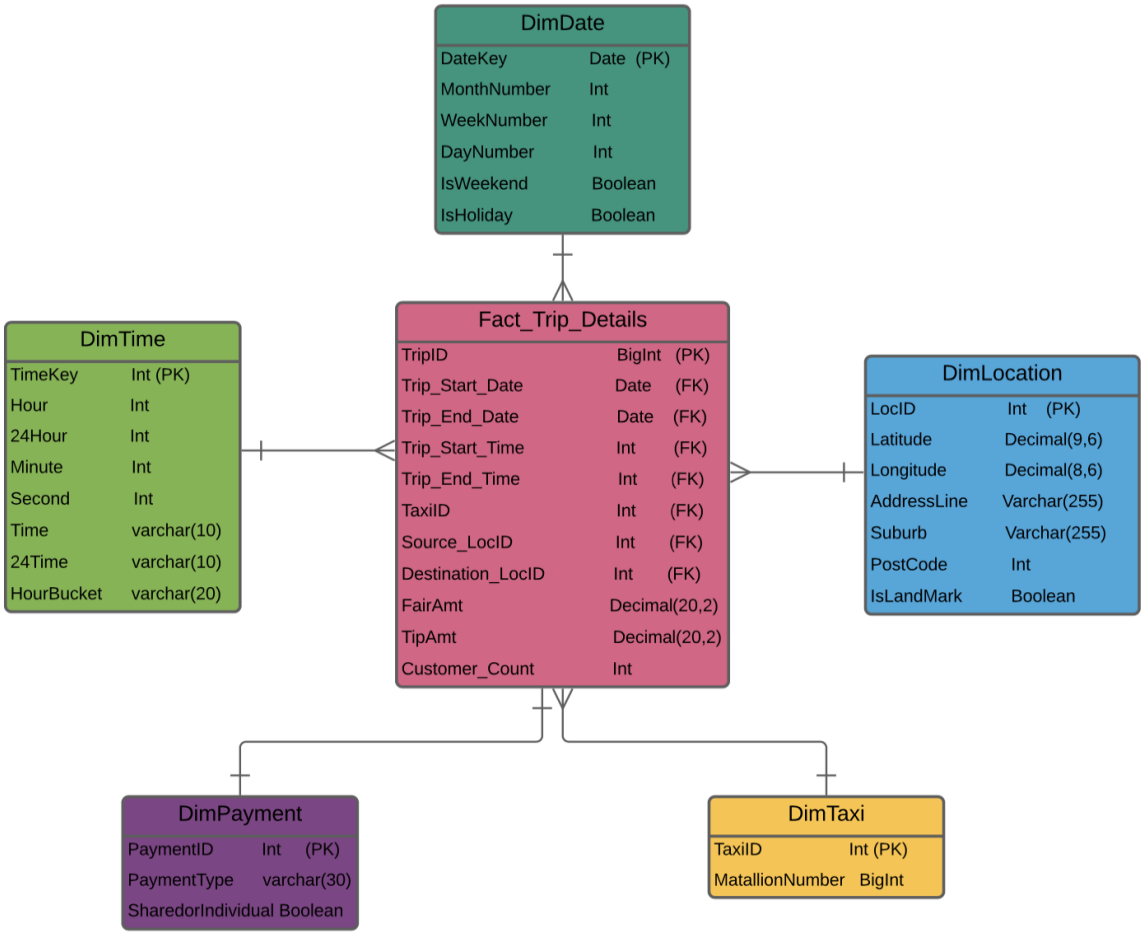
▼ Design for high availability despite the system with huge volume

▼ Partitioning and Delta Processing whenever possible.

Conceptual Data Model



Logical Data Model



Tables

The above two diagrams shows typical flow of a dimensional model of this project. On the first, is a group of tables describing the things about the trip & locations and on the second is a detailed list of tables with Attributes.

Unique Identified for each table is designed to hold business values and the Fact is at the grain of every dimension key: for example, every time for a customer travel, we'll populate a record which provides information at every Location,Date,Time,Payment and also at every driver detail

- List of Tables :

- DimLocation
- DimDate
- DimPayment
- DimTaxi
- FactNYCTrip
- AggNYC_Trip

ETL Processes

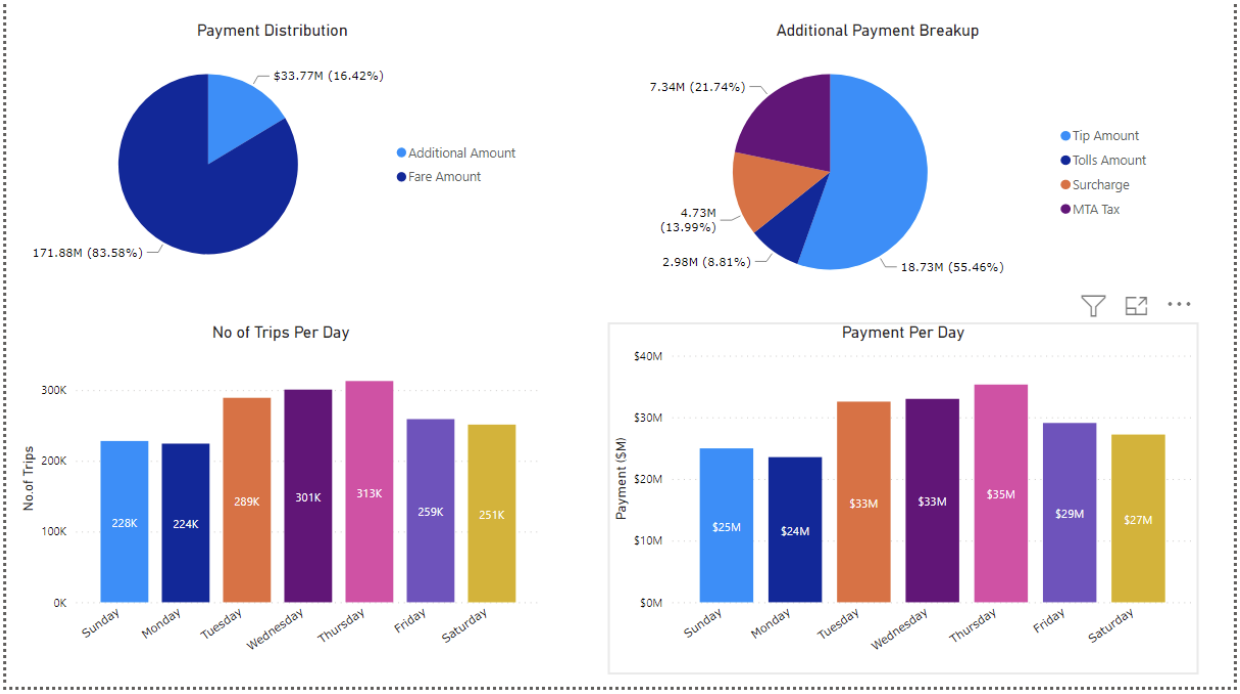
1. Process has been set to load the data in incremental batches (now by date)
2. Partitioning by splitting data based on the date.
3. Currently it has been developed to set for SCD Load Type 1. But it should be enhanced to hold historical value of data for dimensions and append based for transactions.
4. Aggregated and summarised tables are built for specific reporting purposes.

Reporting Layer

▼ Currently a PowerBI reporting model has been created from the aggregated table. As of now, it has been set for every medallion, pickup date, dropoff date, payment type and rate code so the further level of analysis can be done.

▼ Cube Model & Calculations can be built on the PowerBI model using DAX or Microsoft SQL Server Analysis Services.

Sample PowerBI Visuals



SQL Scripting

```
USE [Fact]
GO

DECLARE
@vPICKUP_DATE DATE ,
@vCOUNT INT = 31 ,
@vINDEX INT = 1 ;

DECLARE @tpickup_date Table (
 DATEID INT IDENTITY(1,1) ,
 PICKUP_DATE DATE
)

INSERT INTO @tpickup_date
SELECT [FullDateAlternateKey]
FROM [Fact].[ELULA].[DimDate]
where datekey between 20130101 and 20130131

WHILE( @vINDEX <=  @vCOUNT )
BEGIN

  SELECT @vPICKUP_DATE = PICKUP_DATE FROM @tpickup_date WHERE DATEID = @vINDEX

  INSERT INTO [ELULA].[Fact_NYC_Trip]
         ([medallion]
         ,[hack_license]
         ,[pickup_longitude]
         ,[pickup_latitude]
         ,[dropoff_longitude]
         ,[dropoff_latitude]
         ,[pickup_date]
         ,[dropoff_date]
         ,[pickup_time]
```

```
         ,[dropoff_Time]
         ,[payment_type]
         ,[rate_code]
         ,[passenger_count]
         ,[trip_time_in_secs]
         ,[trip_distance]
         ,[fare_amount]
         ,[surcharge]
         ,[mta_tax]
         ,[tip_amount]
         ,[tolls_amount]
         ,[total_amount])
     SELECT
        A.[medallion]
       ,A.[hack_license]
       ,A.[pickup_longitude]
       ,A.[pickup_latitude]
       ,A.[dropoff_longitude]
       ,A.[dropoff_latitude]
      ,CONVERT(DATE,SUBSTRING(a.[pickup_datetime],1,CHARINDEX(' ',a.[pickup_datetime])),23) AS [pickup_date]
      ,CONVERT(DATE,SUBSTRING(a.[dropoff_datetime],1,CHARINDEX(' ',a.[dropoff_datetime])),23) AS [dropoff_date]
      ,CONVERT(varchar,SUBSTRING(a.[pickup_datetime],CHARINDEX(' ',a.[pickup_datetime]),108)) AS [pickup_time]
      ,CONVERT(varchar,SUBSTRING(a.[dropoff_datetime],CHARINDEX(' ',a.[dropoff_datetime]),108)) AS [dropoff_Time]
      ,b.[payment_type]
      ,a.[rate_code]
      ,cast([passenger_count] as int)
      ,cast([trip_time_in_secs] as int)
      ,cast([trip_distance] as decimal(5,2))
      ,cast([fare_amount] as decimal(20,2))
      ,cast([surcharge] as decimal(20,2))
      ,cast([mta_tax] as decimal(20,2))
      ,cast([tip_amount] as decimal(20,2))
      ,cast([tolls_amount] as decimal(20,2))
      ,cast([total_amount] as decimal(20,2))
    from [dbo].[trip_data_1] A
    INNER JOIN [dbo].[trip_fare_1] B
    ON A.[medallion] = B.[medallion]
    AND A.[hack_license] = B.[hack_license]
    AND A.[pickup_datetime] = B.[pickup_datetime]
    WHERE CONVERT(DATE,SUBSTRING(A.[pickup_datetime],1,CHARINDEX(' ',A.[pickup_datetime])),23) = @vPICKUP_DATE

  SET @vINDEX = @vINDEX + 1
END

GO
```

GitHub Repository

https://github.com/priyavija2020/NYCTrip.git

References

https://chih-ling-hsu.github.io/2018/05/14/NYC

https://docs.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver15&tabs=ssms