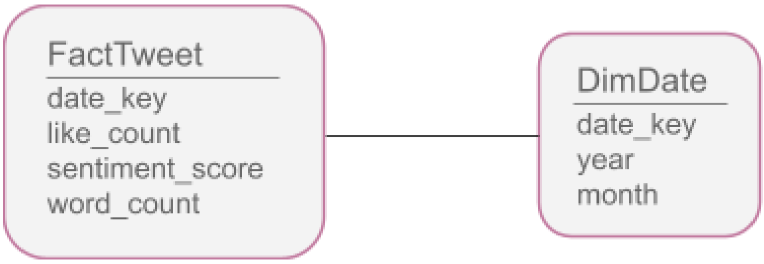
# Assignment 1

## 1. Modelling

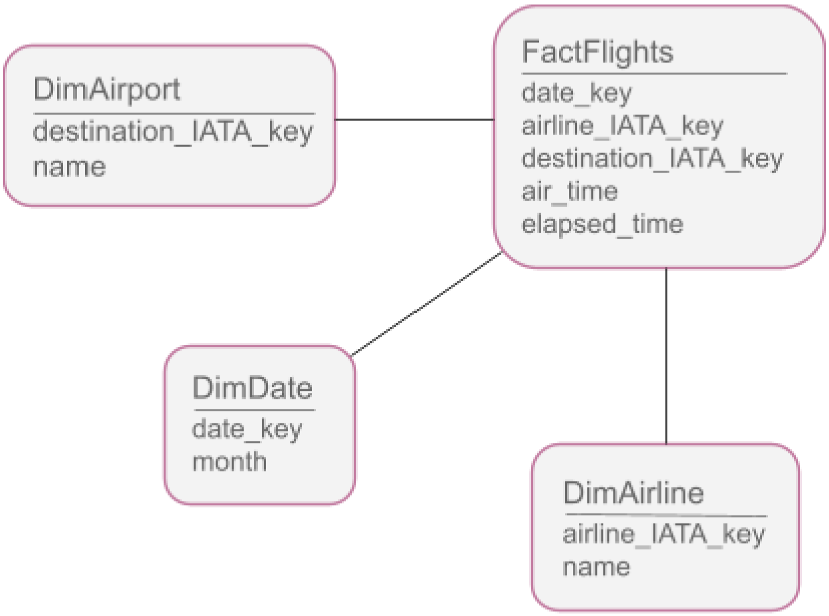
Star Schema of Elon Musk's tweet, I have chosen to leave out all unnecessary attributes from the dataset. For our analysis we need the like count, sentiment score, and word count, as long as a time dimension to filter on dates.
The only relevant concept hierarchy is: *All -> Year -> Month*, for dates.

### Star Schama regarding Report 1 - 4: Elon Tweets



Star schema for Flight data, all unnecessary information has been stripped. For the needed reports we only need air time, elapsed time, destination airport, and airline as relevant measures. Location has not been deemed relevant for this task and therefore has not received a dimension. We will only be looking at months within the dates, thus a hierarchy is irrelevant

### Star Schema regarding report 5 - 8: Flights



## 2. OLAP Operations

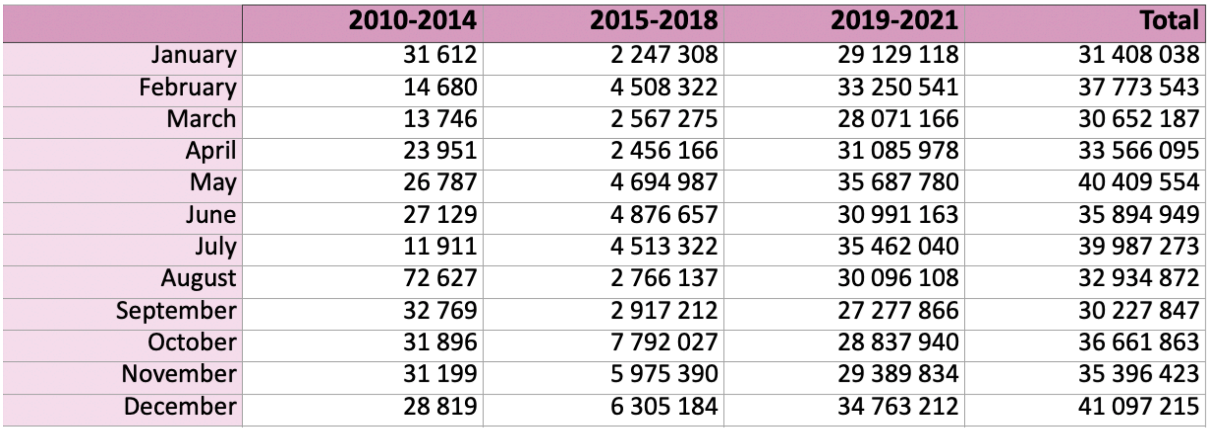
For report one, two, and three, we will need the drill-down operation. Moving from *All*, to *Year* for report two and three, whilst drilling down further to *Month* for report one.

## 3. Implementation of the Cube

**Report 1**: The following query outputs our results:

```
select [Measures].[Like Count] on columns
   [Date].[Month] on rows
   from [Tweets Cube]
```

The resulting table can be collapsed as such, grouping months together. As the data does not cover the whole of April 2021, nor any following months, I have decided to add the average number of likes among January, February, and March from 2021 to April 2021 and the following months of that year. This average amount is 18.383.851.



| | 2010-2014 | 2015-2018 | 2019-2021 | Total |
|---:|---:|---:|---:|---:|
| January | 31 612 | 2 247 308 | 29 129 118 | 31 408 038 |
| February | 14 680 | 4 508 322 | 33 250 541 | 37 773 543 |
| March | 13 746 | 2 567 275 | 28 071 166 | 30 652 187 |
| April | 23 951 | 2 456 166 | 31 085 978 | 33 566 095 |
| May | 26 787 | 4 694 987 | 35 687 780 | 40 409 554 |
| June | 27 129 | 4 876 657 | 30 991 163 | 35 894 949 |
| July | 11 911 | 4 513 322 | 35 462 040 | 39 987 273 |
| August | 72 627 | 2 766 137 | 30 096 108 | 32 934 872 |
| September | 32 769 | 2 917 212 | 27 277 866 | 30 227 847 |
| October | 31 896 | 7 792 027 | 28 837 940 | 36 661 863 |
| November | 31 199 | 5 975 390 | 29 389 834 | 35 396 423 |
| December | 28 819 | 6 305 184 | 34 763 212 | 41 097 215 |

We can then see –assuming the average added is an accurate assumption– that Elon usually gets more likes in **May**, and in **December**. The month with the most likes in total is **February 2021**.

**Report 2**: The following query outputs the tweet count for each year:

```
select [Measures].[Tweet Count] on columns
    [Date].[Year] on rows
    from [Tweets Cube]
```

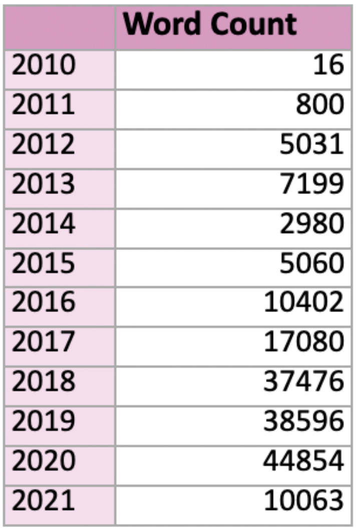
The average number of tweets over ALL years is **1047** for each year, this however includes an outlier being 2010, where there was only one tweet made. If we exclude this the average number over all years is **1142**.

Excluding the year from our query would give us the total number of tweets –which could be divided by the amount of years to achieve an average–, therefore I have opted to include it to remove the outlier that is 2010.

**Report 3**: The following query gives us the word count for each year:

```
select [Measures].[Word Count] on columns
    [Date].[Year] on rows
    from [Tweets Cube]
```

Resulting in the following output:



| | Word Count |
|---|---:|
| 2010 | 16 |
| 2011 | 800 |
| 2012 | 5031 |
| 2013 | 7199 |
| 2014 | 2980 |
| 2015 | 5060 |
| 2016 | 10402 |
| 2017 | 17080 |
| 2018 | 37476 |
| 2019 | 38596 |
| 2020 | 44854 |
| 2021 | 10063 |

**Report 4**: The following query outputs the average sentiment score per month:

```
select [Measures].[Average Sentiment] on columns
    [Date].[Month] on rows
    from [Tweets Cube]
```

**Report 5**: The following query gives us the flight with the longest air time:

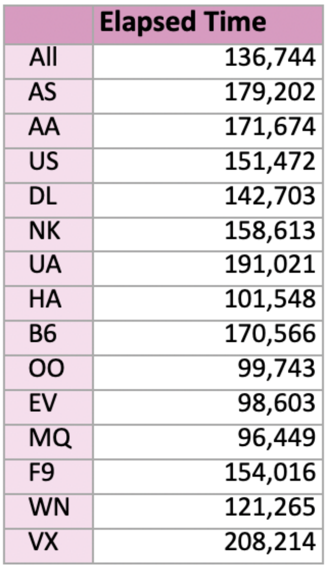
```
select [Measures].[Air Time] on columns
   from [FlightsCube]
```

This gives us flight **51**, tail number **N375HA**, from JFK to HNL.

**Report 6**: The following query gives us the average elapsed time per airline:

```
select [Measures].[Elapsed Time] on columns
   [DimAirline] on rows
   from [FlightsCube]
```

With the following result:

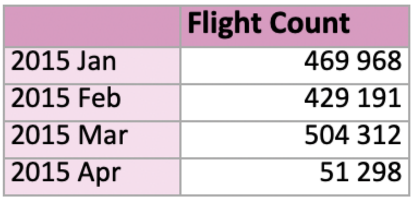


| | Elapsed Time |
|---|---:|
| All | 136,744 |
| AS | 179,202 |
| AA | 171,674 |
| US | 151,472 |
| DL | 142,703 |
| NK | 158,613 |
| UA | 191,021 |
| HA | 101,548 |
| B6 | 170,566 |
| OO | 99,743 |
| EV | 98,603 |
| MQ | 96,449 |
| F9 | 154,016 |
| WN | 121,265 |
| VX | 208,214 |

**Report 7**: The following query gives us the number of flights per month:

```
select [Measures].[Flight Count] on columns
   [Date].[Month] on rows
   from [FlightsCube]
```

Giving us the following table, as we can see February had **426.191** flights.

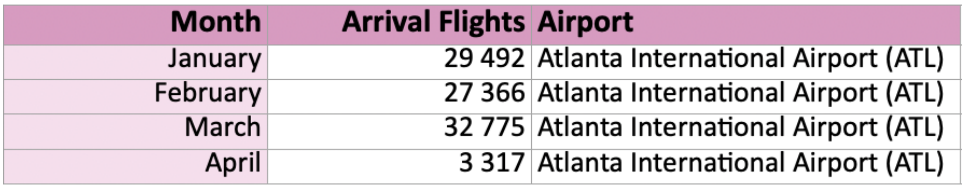


| | Flight Count |
|---|---:|
| 2015 Jan | 469 968 |
| 2015 Feb | 429 191 |
| 2015 Mar | 504 312 |
| 2015 Apr | 51 298 |

**Report 8**: Using the following query we get a spreadsheet we can further use max functions on it to find the airport with the most arrival flights:

```
select [Measures].[Flight Count] on columns
   {[Date].[Month] * [DimAirport]} on rows
   from [FlightsCube]
```

We then get the following results after a spreadsheet max operation:



| Month | Arrival Flights | Airport |
|---:|---:|---|
| January | 29 492 | Atlanta International Airport (ATL) |
| February | 27 366 | Atlanta International Airport (ATL) |
| March | 32 775 | Atlanta International Airport (ATL) |
| April | 3 317 | Atlanta International Airport (ATL) |