Data Warehouse Design

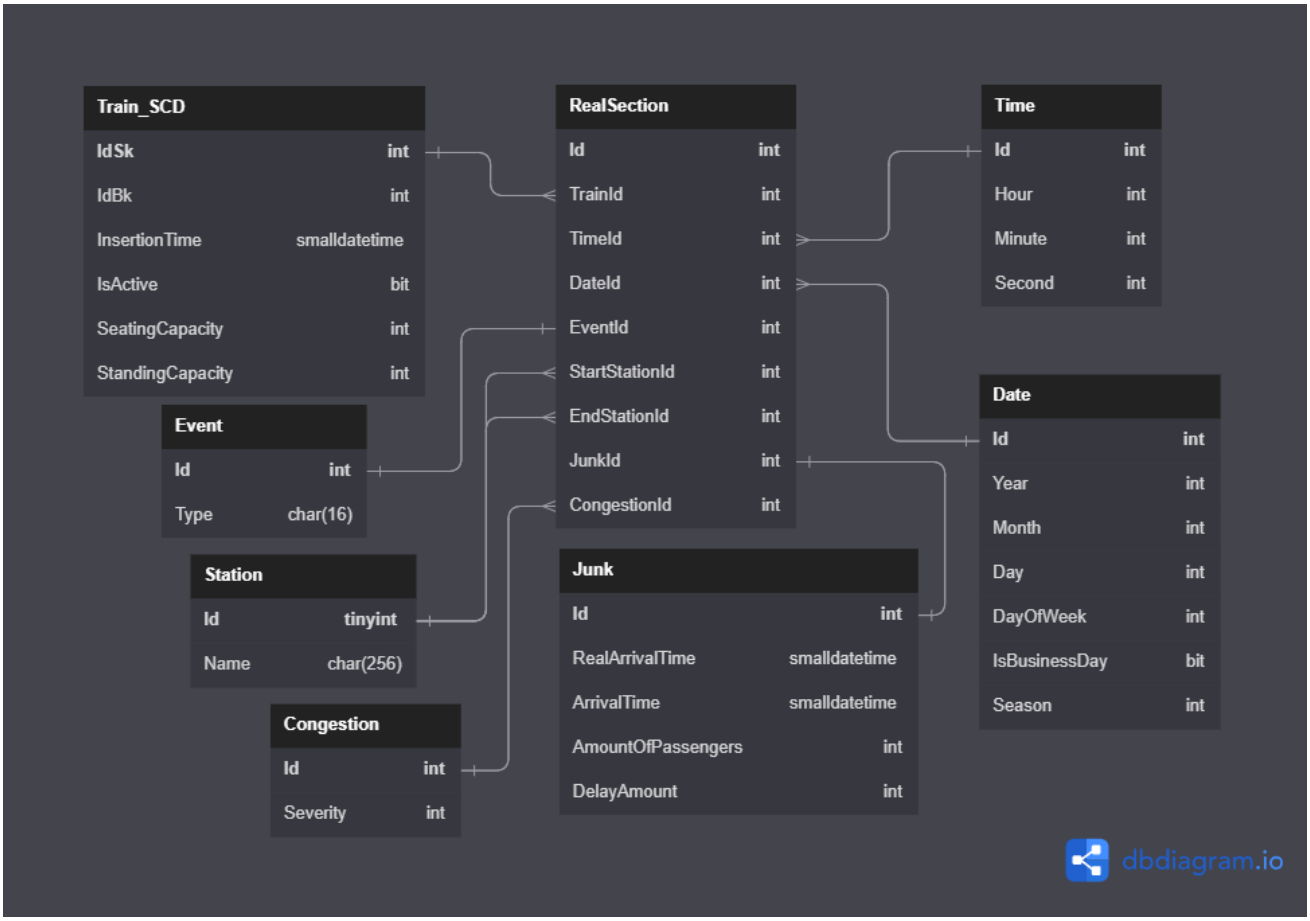

Table Name	Atrribute	Atrribute Type	Description
RealSection (Fact Table)	One tuple describes one fact of the Real Section table.		
	Id	int	PK
	TrainId	int	FK Train
	TimeId	int	FK Time
	DateId	int	FK Date
	EventId	int	FK Event
	StartStationId	int	FK Station Start station of section.
	EndStationId	int	FK Station End station of section.
	JunkId	int	FK Junk
	CongestionId		FK Congestion
Time (Dimension Table)	One tuple describes one hour.		
	Id	int	PK and SK
	Hour	int	Hour. Allowed values from 0 – 23.
	Minute	int	Minute. Allowed vaules from 0-59.
	Second	int	Second. Allowed values from 0-59.
Train (SCD) (Dimension Table)	One tuple describes one train.		
	Id_SK	int	PK and SK
	Id_BK	int	BK
	SeatingCapacity	int	Number of seats available on the train.
	StandingCapacity	int	Number of standing places available on the train.

	IsActive	bit	Allows for scd implementation
	InsertionTime	smalldatetime	Timestamp of insertion to database
Date (Dimension Table)	One tuple describes one date.		
	Id	int	PK and SK
	Year	4 digits	Year
	Month	int	Month number from 0-11, e.g., 0 – January 1 - February
	Day	int	Day (e.g., 25)
	DayOfWeek	int (0;6)	Day of week. Allowed values: 0 (Monday), 1(Tuesday), 2(Wednesday), 3(Thursday), 4(Friday), 5(Saturday) and 6(Sunday)
	IsBusinessDay	bit	Checking if the specific day is a business day.
	Season	int (0;3)	Season. Allowed values: 0(Winter), 1(Spring), 2(Summer), 3(Autumn)
Station	One tuple describes one station.		
	Id	int	PK and SK
	Name	char (256)	Name of the station.
Event	One tuple describes one event.		
	Id	int	PK and SK
	Type	char (256)	There can be some unexpected events during the route: including delays, technical problems with trains, accidents, etc.
Junk (Dimension Table)	The tuples correspond to "all" possible combinations of values for RealArrivalTime, ArrivalTime, DelayAmount and AmountOfPassengers.		

	Id	int	PK and SK
	RealArrivalTime	smalldatetime	Timestamp of actual arrival at the start of the section.
	ArrivalTime	smalldatetime	Timestamp of scheduled arrival at the start of the section.
	DelayAmount	time	Specific amount of time which presents the delay of train.
	AmountOfPassengers	int	Amount of passengers riding on a given section on given train etc
Congestion (Dimension Table)	One tuple describes one congestion.		
	Id	int	PK
	Severity	int	Type of congestion. Allowed values: 0(Light), 1(Medium), 2(Heavy)

Dimensional Model

Fact definitions:

Fact 1 Real Section fact: stores information about number of passengers, train, time, date and events that happened on the section.

Fact table: RealSection

Granularity:

- Train: Holds information about train riding on the section

- Station: Holds station id and name
- Time: Holds time in hours and minutes
- Date: Holds detailed date information
- Congestion: Holds grouped information about amount of people on given section

Measures and aggregation functions:

Number of Passengers- SUM(AmountOfPassengers)

Congestion – AmountOfPassengers - (Train.SeatingCapacity + Train.StandingCapacity)

Throughput – Count(Train between hours x and y) * (Train.SeatingCapacity + Train.StandingCapacity)/1hour

Amount of delay – RealSection.DelayAmount

Dimension definitions:

Dimensions for Fact 1 Real Section fact:

Dimension/Dimension Attribute	**Table/Column**	**Type**
TIME	Time	Dimension
TIME HOUR	Time.Hour	Dimension attribute

TIME MINUTE	Time.Minute	Dimension attribute
TIME SECOND	Time.Second	Dimension attribute
DATE	Date	Dimension
DATE YEAR	Date.Year	Dimension attribute
DATE MONTH	Date.Month	Dimension attribute
DATE DAY	Date.Day	Dimension attribute
DATE DAYOFWEEK	Date.DayOfWeek	Dimension attribute
DATE ISBUSINESSDAY	Date.IsBusinessDay	Dimension attribute
DATE SEASON	Date.Season	Dimension attribute
TRAIN	Train	Dimension
TRAIN STANDING CAPACITY	Train.StandingCapacity	Dimension attribute
TRAIN SEATING CAPACITY	Train.SeatingCapacity	Dimension attribute
TRAIN INSERTION TIME	Train.InsertionTime	Dimension attribute
TRAIN ISACTIVE	Train.IsActive	Dimension attribute
START STATION	Station	Dimension
START STATION NAME	Station.Name	Dimension attribute
END STATION	Station	Dimension
END STATION NAME	Station.Name	Dimension attribute

JUNK	Junk	Dimension
JUNK REAL ARRIVAL TIME	Junk.RealArrivalTime	Dimension attribute
JUNK ARRIVAL TIME	Junk.ArrivalTime	Dimension attribute
JUNK AMOUNT OF PASSENGERS	Junk.AmountOfPassengers	Dimension attribute
JUNK DELAY AMOUNT	Junk.DelayAmount	Dimension attribute

Dimension Name	Dim1Passengers	Dim2Severity	Interval	Expression
REALSECTION CONGESTION HIERARCHY	ALL AmountOfPassengers	Light	< 1300 passengers	AmountOfPassengers <= 1300
	ALL AmountOfPassengers	Medium	1301 – 1600 passengers	AmountOfPassengers between 1301 and 1600
	ALL AmountOfPassengers	Heavy	> 1600 passengers	AmountOfPassengers > 6800

Checking the feasibility of queries based on the multidimensional model:

1. What is the peak hourly throughput on Wednesdays?

Measure: Throughput

Dimension:

- Time (Hour)
- Date (DayOfWeek)

2. What is the size of a deviation from median amount of people in each train, caused by a delayed train.

Measure: Number of Passengers

Dimension:

- ❖ Junk (DelayAmount)

3. What is the average delay of trains on Mondays?

Measure: Amount of Delay

Dimension:

- ❖ Date (DayOfWeek)
- ❖ Junk (DelayAmount)

4. What is the total capacity on business days between 8am-9am?

Measure: Throughput

Dimension:

- ❖ Date (IsBusinessDay)
- ❖ Time (Hour)
- ❖ Train (SeatingCapacity, StandingCapacity)

5. How many hours does the public loose in the summer due to delays?

Measure: Amount of Delay

Dimension:

- ❖ Date (Season)

6. Which stop is the greatest source of delays on Fridays?

Measure: Amount of Delay

Dimension:

- ❖ Junk (DelayAmount)
- ❖ Date (DayOfWeek)

7. What is the change in congestion caused by an event in calendar spring in comparison to mean demand in spring?

Measure: Number of Passengers

Dimension:

- Event (Id)
- Date (Season)
- Congestion (Severity)

8. What is the median delay caused by an event in calendar winter?

Measure: Amount of Delay

Dimension:

- Event (Id)
- Date (Season)
- Junk (DelayAmount)

9. Is the congestion on Mondays heavy only on parts of the route?

Measure: Congestion

Dimension:

- Date (DayOfWeek)
- Station (Id)
- Congestion (Severity)

10. What type of events has the greatest impact on metro congestion on weekends?

Measure: Congestion

Dimension:

- Event (Type)
- Date (IsBusinessDay)
- Congestion (Severity)

Checking if there are Data in the Data sources needed to fill the Data warehouse

Table Name	Column	Source
Train_SCD	One tuple describes one train.	
	IdSK_Train	Train Id. Surrogate key - generated by the database.
	SeatingCapacity	Number of seats available on the train.
	StandingCapacity	Number of standing places available on the train.
Time	One tuple describes a time moment	
	Id_Time	Time Id. Surrogate key - generated by the database.
	Hour	Hour. Allowed values from 0 – 23.
Date	One tuple describes one date.	
	Id_Date	Date Id. Surrogate key - generated by the database.
	IsBusinessDay	Checks if the specific day is a business day.
	DayOfWeek	Day of week. Allowed values: 0 (Monday), 1(Tuesday), 2(Wednesday), 3(Thursday), 4(Friday), 5(Saturday) and 6(Sunday)
	Season	Season. Allowed values: 0(Winter), 1(Spring), 2(Summer), 3(Autumn)

Event	One tuple describes one event.	
	Id_Event	Event Id. Surrogate key - generated by the database.
	Type	Stores information about Type of the event in string format. Usually causes a delay.
Station	One tuple describes one station.	
	Id_Station	Station Id. Surrogate key - generated by the database.
Congestion	One tuple corresponds to a single congestion category.	
	Id_Congestion	Congestion Id. Surrogate key - generated by the database.
	Congestion_Severity	Type of congestion. Allowed values: 0(Light), 1(Medium), 2(Heavy)
Junk	The tuples correspond to "all" possible combinations of values for RealArrivalTime, ArrivalTime, AmountOfPassengers and DelayAmount and are generated before ETL process.	
	Id_Junk	Junk Id. Surrogate key - generated by the database.
	DelayAmount	Specific amount of time which presents the delay of train.