StarBus - Bus Transport Organization Data Warehouse Design

Business process

The Date warehouse is designed for Bus Journeys Optimisation business process. This process is described in the document *Specification of business processes in "StarBus" Bus Transport Organization.*

Relational Database schema

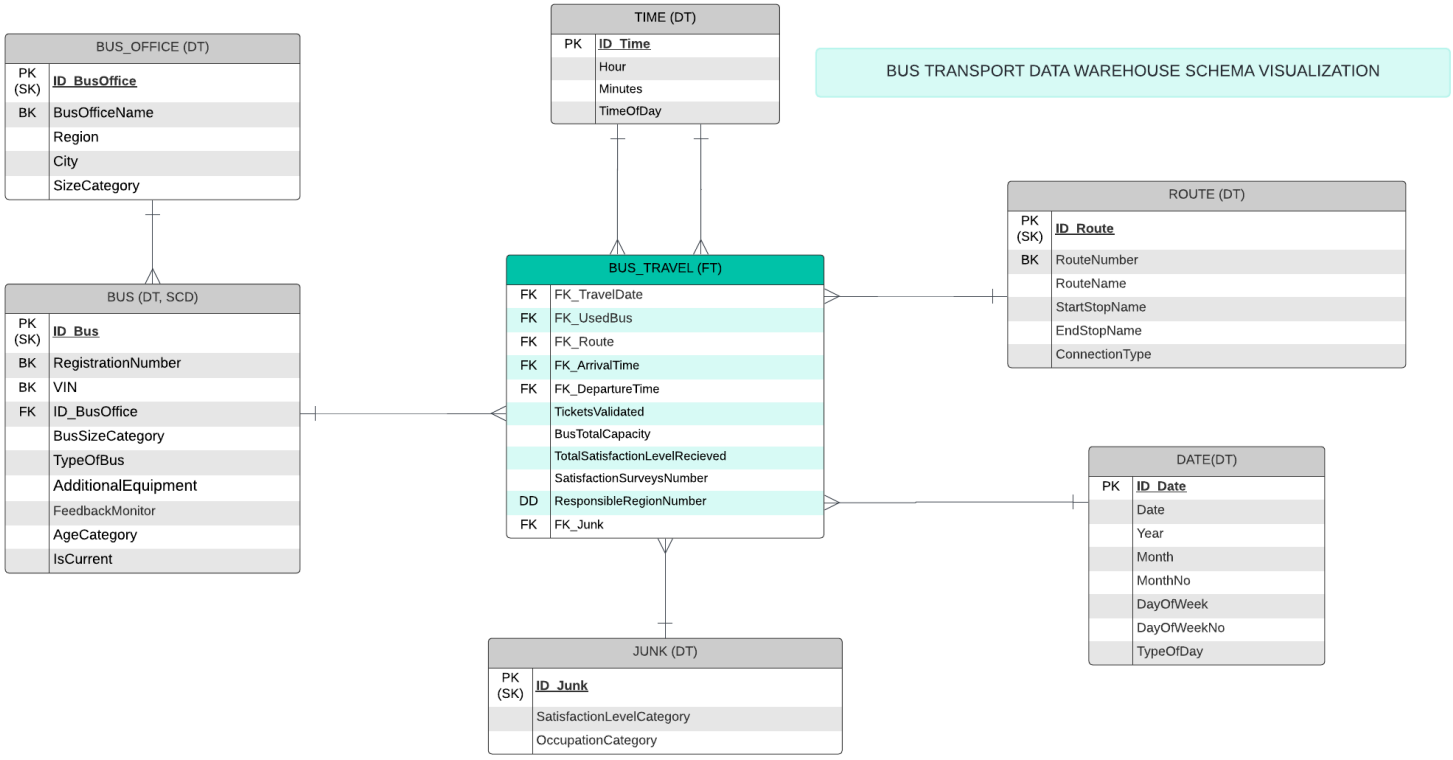

TABLE NAME	ATTRIBUTE	ATTRIBUTE TYPE	DESCRIPTION

BUS_TRAVEL (FACT TABLE)	One tuple describes one fact of a bus travel, specified by used bus, date of travel, route, arrival and departure time		
	ID_TravelDate	Numeric	FK Date The date when the travel took place
	ID_UsedBus	Numeric	FK Bus ID of used bus
	ID_Route	Numeric	FK Route Number of bus route
	ID_ArrivalTime	Numeric	FK Arrival time Time of bus arrival
	ID_DepartureTime	Numeric	FK Departure time Time of bus departure
	TicketsValidated	Numeric	Sum of number of tickets validated for a given travel
	BusTotalCapacity	Numeric	Maximum capacity of a single bus
	TotalSatisfactionLevelReceived	Numeric	Sum of all satisfaction's levels received for a given travel
	SatisfactionSurveysNumber	Numeric	Count of received feedback for a given travel
DD	ResponsibleRegionNumber	Numeric	Number of specified voivodeship of Poland
	ID_Junk	Numeric	FK Junk FK attributes
BUS (DIMENSION TABLE, SCD)	One tuple describes one bus on which the travel was made, with specified registration number, VIN, type of bus, additional equipment and age category		
	ID_Bus		PK (surrogate key)
	RegistrationNumber	Varchar (2-8)	BK, unique polish registration number of the bus (consist of characters and numbers)
	VIN	Varchar (17)	BK, vehicle's unique identification number
	ID_BusOffice	Numeric	Bus office identification number, which indicates office that is responsible

			for given bus, its maintenance and travelled routes
	BusSizeCategory	Varchar (3-15)	Size of bus. Allowed values: small, average, big
	TypeOfBus	Varchar (3-9)	Type of a bus, could be one of following options: *minibus, standard, low floor*
	AdditionalEquipment	Varchar (2-3)	Indicate whether the bus has additional equipment such as a place for a disabled person or air conditioning. Allowed values: “yes” - if bus has any additional equipment, “no” - otherwise
	FeedbackMonitor	Varchar (2-3)	Indicate whether the bus has feedback monitor inside. Allowed values: “yes” - if bus has feedback monitor, “no” - otherwise.
	AgeCategory	Varchar (3-6)	Bus is categorized according to the difference between current year and production year: less than 3 (*new*), between 3 and 10 (*middle*), more than 10 (*old*)
	IsCurrent	Boolean	1 if information is current, otherwise 0 (SCD implementation)
BUS_OFFICE (DIMENSION TABLE)	One tuple describes one bus office by identification number, name, region, city and bus slots		
	ID_BusOffice	Numeric	PK (surrogate key), Bus office identification number, which indicates office that is responsible

			for given bus, its maintenance and travelled routes
	BusOfficeName	Varchar (20 - 40)	BK, Bus Office's Name
	Region	Varchar (5-20)	Single voivodeship of Poland, indicate for what region given bus office is responsible,
	City	Varchar (5-15)	City of Poland, where the office is located
	SizeCategory	Varchar (3-15)	Bus Office size. Allowed values: *small, average, big*
TIME (DIMENSION TABLE)	One tuple describes one hour (independently on date)		
	ID_Time	Numeric	PK
	Hour	Numeric	Hour. Allowed values from 0 – 23
	Minutes	Numeric	Minutes. Allowed values from 0 – 59
	TimeOfDay	Varchar (20)	Time of day (time intervals). Allowed values: *0-8, 9-12, 13-15, 16-20, 21-23*
ROUTE (DIMENSION TABLE)	One tuple describes one connection provided by the organization, routes are identified by route number		
	ID_Route	Numeric	PK (surrogate key)
	RouteNumber	Numeric	BK, unique number of bus route
	RouteName	Varchar (2-30)	Bus route name
	StartStopName	Varchar (5-40)	Name of the start stop of the route
	EndStopName	Varchar (5-40)	Name of the end stop of the route
	ConnectionType	Varchar (7-15)	Indicates whether a given route has at least one of the stops close to airport, Allowed Values:

			Airport (whether at least one of the stops is close to airport), *Non-Airport* (whether none of the stops are close to airport)
DATE (DIMENSION TABLE)	One tuple describe one day.		
	ID_Date	Numeric	PK
	Date	Date	Date
	Year	4 digits	Year
	Month	Varchar (3-10)	Month. Allowed values: *January, February, March, April, May, June, July, August, September, October, November and December*
	MonthNo	Numeric	Month's numeric value
	DayOfWeek	Varchar (6-10)	Day of week. Allowed values: *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday*
	DayOfWeekNo	Numeric	Weekday's numeric value
	TypeOfDay	Varchar (5-15)	Type of day of the week. Allowed values: *working day, weekend, holiday/vacation*
JUNK (DIMENSION TABLE)	The tuples correspond to "all" possible combinations of values for SatisfactionLevelCategory and OccupationCategory		
	ID_Junk	Numeric	PK
	SatisfactionLevelCategory	Varchar (3-6)	SatisfactionLevelCategory Of bus travel (<=3 - *low*, 3<x<=7 - *medium*, 7< - *high*)
	OccupationCategory	Varchar (3-6)	OccupationCategory of the bus travel (<15% - *very low*, <=30% - *low*, <=75% - *medium*, >75% - *high*)

Dimensional Model

Fact definition

Fact 1 Travel Fact: Travel realized by a specified bus with a certain bus occupancy on a specified day and time (scheduled arrival, departure). It is realized according to specified route and supervising region, the one from where bus starts its journey.

Fact Table: Bus_Travel

Granularity:

- a specified responsible region,
- a specified bus maximum capacity
- a specified surveys number received
- a specified total level of satisfaction
- a specified arrival time
- a specified departure time
- a specified date of travel with specified type of day
- a specified scheduled route with start stop and destination/end stop, and its connection category
- a specified bus that is of specified type, size category, at specified AgeCategory, and with specified additional equipment – wheelchair, air-conditioning or feedback monitor.

Measures and aggregation functions:

Number of travel facts – COUNT (1)
Number of tickets validated - SUM(TicketsValidated)
Total number of bus capacity – SUM(BusTotalCapacity)
Number of Satisfactions Received – SUM(SatisfactionSurveysNumber)
Total Satisfaction Level Received - SUM(SatisfactionLevelRecieved)
Number of operating regions – DISTINCT COUNT(ResponsibleRegionNumber)

Average satisfaction level -
Total Satisfaction Level Received/*Number of Satisfactions Received*

Average number of passengers -
Number of tickets validated/*Number of travel facts*

Bus Occupancy Percentage -
Number of tickets validated/*Total number of bus capacity* * 100%

Dimension definitions

Dimensions for Fact 1 Travel fact

DIMENSION/DIMENSION ATTRIBUTE	TABLE/COLUMN	TYPE
REGION NUMBER	Bus_Travel.ResponsibleRegionNumber	Degenerate dimension
BUS	Bus	Dimension
REGISTRATION NUMBER	Bus.RegistrationNumber	Dimension attribute
VIN	Bus.VIN	Dimension attribute
BUS SIZE CATEGORY	Bus.BusSizeCategory	Dimension attribute
TYPE OF BUS	Bus.TypeOfBus	Dimension attribute
ADDITIONAL EQUIPMENT	Bus.AdditionalEquipment	Dimension attribute
FEEDBACK MONITOR	Bus.FeedbackMonitor	Dimension attribute
BUS AGE CATEGORY	Bus.AgeCategory	Dimension attribute
BUS HIERARCHY	• Bus.TypeOfBus • • Bus.AgeCategory • • • Bus.RegistrationNumber • • • • Bus.VIN	Hierarchical dimension
BUS OFFICE	BusOffice	Dimension
BUS OFFICE NAME	BusOffice.Name	Dimension attribute
REGION	BusOffice.Region	Dimension attribute
CITY	BusOffice.City	Dimension attribute
BUS OFFICE SIZE CATEGORY	BusOffice.SizeCategory	Dimension attribute
BUS OFFICE LOCALIZATION HIERARCHY	• BusOffice.Region • • BusOffice.City • • • BusOffice.Name	Hierarchical dimension
TRAVEL DATE	Date	Dimension
TRAVEL YEAR	Date.Year	Dimension attribute
TRAVEL MONTH	Date.Month	Dimension attribute
TRAVEL DAY OF WEEK	Date.DayOfWeek	Dimension attribute
TRAVEL TYPE OF DAY	Date.TypeOfDay	Dimension attribute
TRAVEL DAY OF WEEK HIERARCHY	• Date.Year • • Date.Month • • • Date.DayOfWeek	Hierarchical dimension
TRAVEL TYPE OF DAY HIERARCHY	• Date.Year • • Date.Month • • • Date.TypeOfDay	Hierarchical dimension
TRAVEL DATE HIERARCHY	• Date.Year • • Date.Month • • • Date.Date	Hierarchical dimension

TIME	Time	Dimension
HOUR	Time.Hour	Dimension attribute
MINUTES	Time.Minutes	Dimension attribute
TIME OF DAY	Time.TypeOfDay	
TRAVEL TIME HIERARCHY	• Time.Hour • • Time.Minutes • • • Time.TypeOfDay	Hierarchical dimension
JUNK	Junk	Dimension
SATISFACTION LEVEL CATEGORY	Junk.SatisfactionLevelCategory	Dimension attribute
OCCUPATION CATEGORY	Junk.OccupationCategory	Dimension attribute
ROUTE	Route	Dimension
ROUTE NUMBER	Route.RouteNumber	Dimension attribute
ROUTE NAME	Route.RouteName	Dimension attribute
START STOP NAME	Route.StartStopName	Dimension attribute
END STOP NAME	Route.EndStopName	Dimension attribute
CONNECTION TYPE	Route.ConnectionType	Dimension attribute

Checking the feasibility of queries based on the multidimensional model

1. Compare the average number of passengers per route on working days and weekends in current and previous month.

Measure: Average number of passengers

Dimension: Junk (dimension attribute: OccupationCategory)

Dimension: Travel date (dimension attributes: Travel TypeOfDay)

2. From which region did the bus office operate the most bus travel?

Measure: Number of travel facts

Dimension: Bus Office (dimension attributes: Region)

3. How airport connections have an impact on bus occupancy?

Measure: Bus Occupancy Percentage

Dimension: Route (dimension attributes: ConnectionType)

Dimension: Junk (dimension attributes: OccupationCategory)

4. Compare the average satisfaction level for all routes in the current month to those in the previous month.

Measures: Average satisfaction level

Dimension: Route (dimension attributes: RouteNumber, RouteName)

Dimension: Travel date (dimension attributes: Travel month)

5. Whether passenger satisfaction level decreases when buses exceed a certain occupancy threshold? (<15%, <30%, 30-75%, >75%)

Measures: Bus Occupancy Percentage

Dimension: Junk (dimension attributes: OccupationCategory)

6. Do certain routes are associated with more frequent passenger feedback?

Measures: Number of Satisfactions Received

Dimension: Route (dimension attributes: RouteNumber, RouteName)

7. Are passengers more likely to complete surveys when travelling on a bus with a feedback monitor installed?

Measures: Number of Satisfactions Received

Dimension: Bus (dimension attributes: Feedback Monitor)

8. Which route has the lowest and highest average satisfaction level in the survey?

Measures: Average satisfaction level

Dimension: Route (dimension attributes: RouteNumber, RouteName)

Dimension: Junk (dimension attributes: SatisfactionLevelCategory)

9. Do additional amenities such as air conditioning or wheelchair access affect satisfaction levels?

Measures: Average satisfaction level

Dimension: Bus (dimension attributes: AdditionalEquipment)

Dimension: Junk (dimension attributes: SatisfactionLevelCategory)

10. Are old buses (Production Year < 2010) receiving lower ratings?

Measures: Average satisfaction level

Dimension: Bus (dimension attributes: AgeCategory)

Dimension: Junk (dimension attributes: SatisfactionLevelCategory)

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

TABLE NAME	COLUMN	SOURCE
BUS_TRAVEL	One tuple describes one fact of a bus travel.	
	ID_TravelDate	Travel Date Id. Foreign key from dimension table. Based on IssueDate stored in Travel table in RoutesTraveller source.
	ID_UsedBus	Travel Bus Id. Foreign key from dimension table. Based on Registration Number and its maximum capacity. Based on foreign key FK_Bus stored in table Travel in RoutesTraveller source and sheet 2 in Service_Office EXCEL.
	ID_Route	Route Id. Foreign key from dimension table. Based on Route Number and Name. Based on foreign key FK_Route stored in table Travel in RoutesTraveller source.
	ID_ArrivalTime	Travel Arrival Time Id. Foreign key from dimension table. Based on arrival time value stored in Schedule table and foreign key FK_Schedule in Travel table in RoutesTraveller source.
	ID_DepartureTime	Travel Departure Time Id. Foreign key from dimension table. Based on departure time value stored in Schedule table and foreign key FK_Schedule in Travel table in RoutesTraveller source.

	TicketsValidated	Sum of all tickets validated on given travel, value taken from sum of FK_Ticket in Validation table in RoutesTraveller source for single FK_Travel.
	BusTotalCapacity	Bus maximum possible capacity for single bus travel. It is a sum of seats (column G) and standing places (column H) in sheet 2 in Service_Office EXCEL.
	TotalSatisfactionLevelRecieved	Sum of all satisfaction's levels received for given travel, value based on SatisfactionLevel and FK_Travel columns in Feedback table in RoutesTraveller source.
	SatisfactionSurveysNumber	Count existence of feedback for given travel. Total number of surveys left. Based on FK_Travel in Feedback table in RoutesTraveller source.
DD	ResponsibleRegion	Region taken from sheet 1 in Service_Office EXCEL with number from 1 – 5 assigned alphabetically. Based on region in which service office is placed (value in column F – Region) that is responsible for bus assigned to specific travel.
	ID_Junk	Junk Id. Foreign key from dimension table. Based on SatisfactionLevelCategory and OccupationCategory from Travel table in RoutesTraveller source.
BUS	One tuple describes a bus in which the travel was made	
	ID_Bus	Bus Id. Surrogate key - generated by database.
	RegistrationNumber	Business key taken from RegistrationNumber from Bus table in RoutesTraveller source.
	VIN	Business key taken from VIN from Bus table in RoutesTraveller source.
	ID_BusOffice	BusOffice Id. Foreign key from dimension table. Value based on data from sheet 2 in Service_Office EXCEL.

	BusSizeCategory	Bus maximum possible capacity assigned to category. Based on a sum of seats (column G) and standing places (column H) in sheet 2 in Service_Office EXCEL. Allowed values: *small* - for <=8 places, *average* - for <=17 places, *big* - for >17 places.
	TypeOfBus	Type of a bus. Allowed values: minibus, standard, low floor. Based on the value stored in sheet 2 in Service_Office EXCEL.
	AdditionalEquipment	AdditionalEquipment in a bus. Based on the values stored in columns I, J (wheelchair, air conditioning) from sheet 2 in Service_Office EXCEL. Allowed values: "*yes*" - if bus has any additional equipment, "*no*" – otherwise.
	FeedbackMonitor	Feedback monitor inside a bus stored in column K – Feedback Monitor in sheet 2 in Service_Office EXCEL. Allowed values: "*yes*" - if bus has feedback monitor, "*no*" – otherwise.
	AgeCategory	Bus Age Category. Allowed values: less than 3 (*new*), between 3 and 10 (*middle*), more than 10 (*old*). Calculated based on current year and column F – Production Year in sheet 2 in Service_Office EXCEL. (SCD2 implementation).
	IsCurrent	"1" if information is current, otherwise "0" (SCD implementation).
BUS_OFFICE	One tuple describes one bus office with specified name, region, city and number of bus slots.	
	ID_BusOffice	Bus Office Id. Surrogate key – generated by database.
	BusOfficeName	Bus Office Name. Business key taken from Name (column B) from sheet 1 in Service_Office EXCEL.

	Region	Voivodeship of Poland where given bus office is located, used to indicate for what region given bus office is responsible. Based on column F in sheet 1 from Service_Office EXCEL.
	City	City of Poland, where the office is located. Based on values from column E from sheet 1 in Service_Office EXCEL.
	SizeCategory	Category assigned to number of bus slots in given office, value based on number taken from column H from sheet 1 in Service_Office EXCEL. Allowed values: *small* – for 8 bus slots, *average* – for 9 bus slots, *big* – for 10 bus slots
TIME	One tuple describes one hour (independently on date) All the data in this table are generated tuple by tuple based on clock, before ETL process.	
ROUTE	One tuple describes one connection provided by the organization, routes are identified by route number	
	ID_Route	Route Id. Surrogate key – generated by database.
	RouteNumber	Business key taken from RouteNumber from Route table in RoutesTraveller source.
	RouteName	Route Name. Value taken from RouteName column in Route table in RoutesTraveller source.
	StartStopName	Start Stop Name. Based on value in StartStopID column in Route table in RoutesTraveller source and corresponding name value from column B, sheet 1 in Bus_Stops EXCEL.
	EndStopName	End Stop Name. Based on value in StartStopID column in Route table in RoutesTraveller source and corresponding name value from column B, sheet 1 in Bus_Stops EXCEL.
	ConnectionType	Type of a connection. Based on StartStopID and EndStopID “Airport”

		properties from column F, sheet 1 in Bus_Stops EXCEL. Allowed Values: - *Airport* (whether at least one of the stops is close to airport), - *Non-Airport* (whether none of the stops are close to airport).
DATE	One tuple describe one day. All the data in this table are generated tuple by tuple based on any calendar, before ETL process	
JUNK	The tuples correspond to "all" possible combinations of values for SatisfactionLevelCategory and OccupationCategory and are generated before ETL process.	
	ID_Junk	Junk Id. Surrogate key - generated by database.
	SatisfactionLevelCategory	Satisfaction Level Category. Based on SatisfactionLevelReceived column in Bus_Travel table in datawarehouse. Allowed values: Level <=3 - *low*, Level <=7 – *medium*, Level <=10 – *high.*
	OccupationCategory	Occupation Category. Based on *BusOccupatioRatio* – column in Bus_Travel table in data warehouse, multiplied by 100%. Allowed values: Occupation <=15% - *Very low* (possibility of non-useful connection) Occupation <=30% - *low* Occupation <=75% - *medium* Occupation >75% - *high*