Requirements specification for Bus Journeys Optimisation business process

1. General description of business process

a. A general description of the business process and a description of the performance metrics generated by this process, possible current analytical problems.

The process begins with passengers awaiting the arrival of buses at bus stops. Bus might arrive according to the exact schedule or with some delay. As the bus approaches, each passenger must scan their ticket, which is done electronically using ticket-scanning equipment and QR code (the bought ticket). System collects information about validated tickets at start station and after bus departure update it to the system. System assumes that passengers are leaving the bus at their destination stop assigned to their ticket. On the app, in some of the buses or via link sent by email they can complete anonymous satisfaction survey about their travel. Delays, bus occupancy, and overall service quality might have an impact. System stores those satisfactions levels per each travelled route.

The monthly increase by at least 0,5% in average satisfaction levels of the 'travelled' routes compared to previous month.
The average bus occupancy at the end of 2024-year will be not less than 15% of total bus capacity.

**We assume that desired bus occupancy will be met for at least 80% of all travelled routed.*

b. Typical questions

What are the most overloaded routes?
How do overload of buses affect passenger satisfaction?
On which days are buses most crowded weekdays, weekends or holidays?
Which routes receive the highest/the lowest satisfaction ratings from passengers?
Identify routes with the lowest bus occupancy (<15% of total bus capacity).
From which region did the bus office operate the most bus travel?
Compare the average number of passengers depending on the type of bus in previous and current month.
Do certain routes are associated with more frequent passenger feedback?
Do additional amenities such as air conditioning or wheelchair access affect satisfaction?
Are old buses (Production Year < 2010) receiving lower ratings?

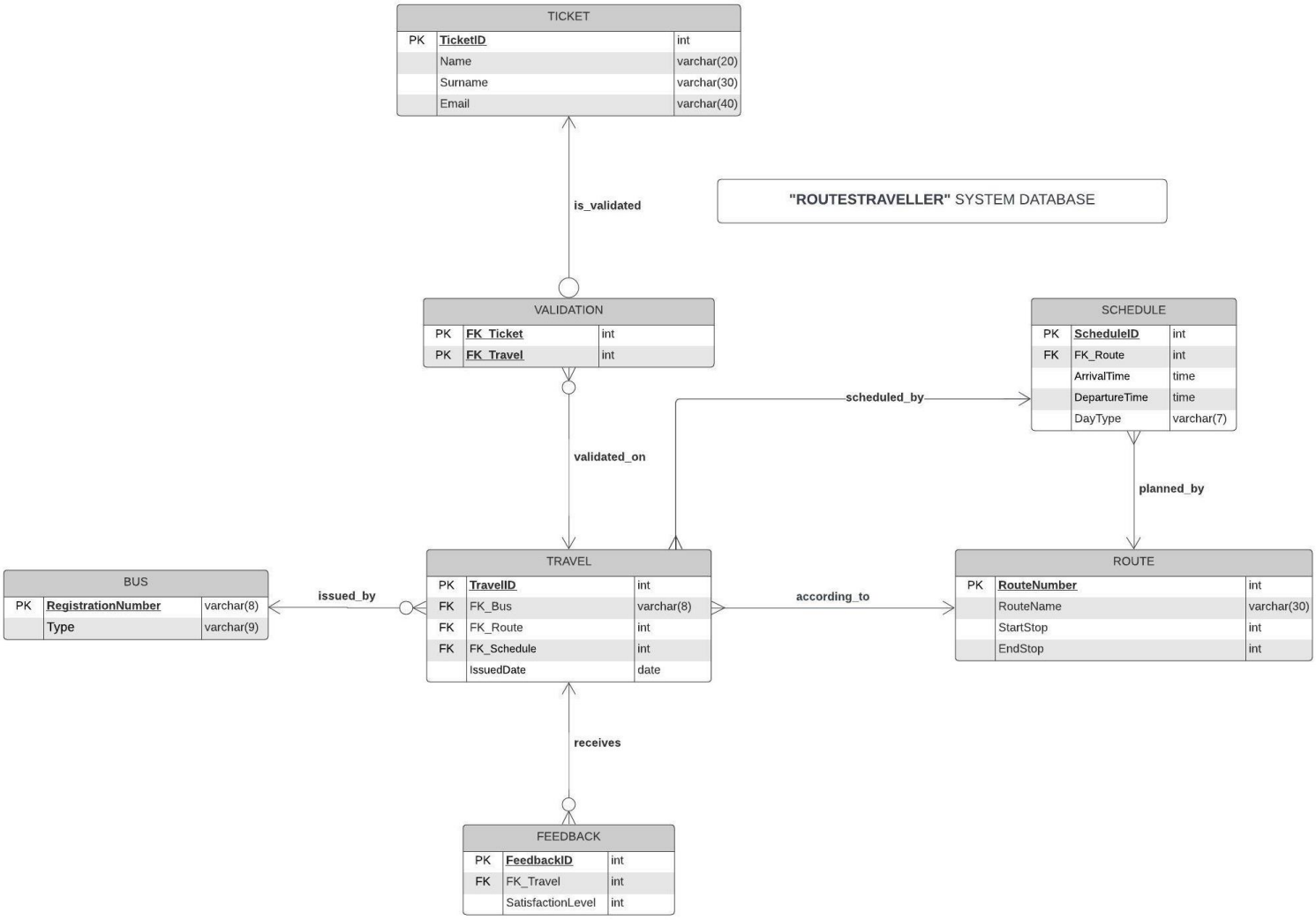
c. Data

Data sources include "RoutesTraveller" system with information about validated tickets, buses and its issued travels on specified routes. It also contains feedback (satisfaction levels) assigned to the travel as well as time-scheduled arrivals, departures at bus stops. Data about buses and their service offices is stored in separate EXCEL file. Moreover, specified information about all bus stops on the routes is also stored in separate EXCEL file.

2. Data sources structures

RoutesTraveller

Entity Relationship Diagram



Entities Description

TABLE NAME	ATTRIBUTE	ATTRIBUTE TYPE	DESCRIPTION
BUS	Single Bus vehicle identified by registration number		
	RegistrationNumber	String – (2-8) characters	PK, unique polish registration number of the bus (consist of characters and numbers)
	Type	String – (3-9) characters	Type of a bus, could be one of following options: minibus, standard, low floor
TICKET	Specific ticket purchased by a passenger for a specific travel starting and finishing at designated stops, identified by ticket number		
	TicketID	int	PK, unique, auto generated starting from 1 with increment 1
	Name	String – (2-20) characters	Passenger's name, only ASII characters (letters), without numbers
	Surname	String – (2-50) characters	Passenger's surname, only ASII characters (letters), without numbers
	Email	String – (4-40) characters	Passenger's email, must include '@'
VALIDATION	Validation of the ticket, during travel		
	FK_Ticket	int	FK, part of PK, relationship **is_validated** ('ticket' **1:0...1** 'validation')
	FK_Travel	int	FK, part of PK, relationship **validated_on** ('travel' **1:0...n** 'validation')
TRAVEL	Specific bus travel identified by TravelID.		
	TravelID	int	PK, unique, auto generated starting from 1 with increment 1
	FK_Bus	String – (2-8) characters	FK, relationship **issued_by** ('bus' **1:0...n** 'travel')
	FK_Route	int	FK, relationship **according_to** ('travel' **1...n:1** 'route')
	FK_Schedule	int	FK, relationship **scheduled_by** ('travel' **1...n:1** 'schedule')
	IssuedDate	date*	Date of bus travel - when bus start its assignment, date of its arrival at the start stop.
FEEDBACK	Anonymous feedback from passengers about a specific travel. Identified by FeedbackID.		

	FeedbackID	int	PK, unique, auto generated starting from 1 with increment 1
	FK_Travel	int	FK, relationship **receives** (‘travel’ **1:0...n** 'feedback’)
	SatisfactionLevel	int (range 1-10)	Passenger satisfaction score in scale from 1 to 10.
SCHEDULE	Bus timetable of specific bus routes identified by ScheduleID. Time departures and arrivals can differ due to DayType value.		
	ScheduleID	int	PK, unique, auto generated starting from 1 with increment 1
	ArrivalTime	time**	Estimated time of bus arrival
	DepartureTime	time**	Estimated time of bus departure
	DayType	String – (3-7) characters	Type of day of the week. Allowed values: weekday, weekend, holiday
	FK_Route	int	FK, relationship **planned_by** (‘route’ **1:1...n** 'schedule’)
ROUTE	Currently connections provided by the organization, routes are identified by route number		
	RouteNumber	int	PK, unique number of route
	RouteName	String – (2-30) characters	Route's name
	StartStop	int	Unique identification number of a bus start stop, relates to *Bus_Stops EXCEL*
	EndStop	int	Unique identification number of a bus end stop, relates to *Bus_Stops EXCEL*

* *date format YYYY-MM-DD* - eight digits separated by '-' sign. Restrictions: 'DD' day number in range 01-31, 'MM' month number in range 01-12, 'YYYY' - year 4 digits.

** time format hh:mm – 6 digits separated by ‘:’, ‘hh’ hour 00-23 number and ‘mm’ minutes in range 00-59.

Relations Description

- ***is_validated*** (‘ticket’ **1:0...1** ‘validation’) - shows which ticket has been validated on specific travel, ticket can be obtained by passenger, but it can be still not validated and therefore it is not assigned to specific travel, each validation always represents only one ticket without optionality.

- ***validated_on*** (‘travel’ **1:0...n** ‘validation’) - shows on which travel specific ticket was validated, each travel might have zero, one or many validations (e.g. there might be routes when no one has bought the ticket and validated it in a bus), each validation relates to exactly one travel, with no optionality.

- ***issued_by*** (‘bus’ **1:0...n** ‘travel’) – shows which bus was assigned to travel, each bus might issue zero, one or many travels (e.g. new buses has 0 completed travels, and older ones have many), each travel must be issued by exactly one bus, with no optionality.

- ***according_to*** (‘travel’ **1...n:1** 'route’) - shows to what specific route given travel relates to, that determines multiple rides according to the same route, each travel is realized according to exactly one route (from bus stop A to bus stop B), each route can have one or many travels on it (there are no routes that have been never travelled by at least one bus, if organization decides to add a new route, it is added with the first travel realized on it, respectively with new schedule).

- ***scheduled_by*** (‘travel’ **1...n:1** 'schedule’) - shows according to what schedule is given travel planned (time arrivals at start stop and end stop), used to distinguish different time of travels realized on the same routes and at the same days, each travel is planned by exactly one schedule (time of arrival on start stop A and start stop B), each schedule can have one or many travels according to it (without optionality, if there are necessary schedule changes, they are updated immediately to the system with the new travel – same situation as with adding new route)

- ***receives*** (‘travel’ **1:0...n** 'feedback’) - shows to which specific travel given passenger’s feedback relates to, each travel can receive might have zero, one or much feedback (e.g. no one has filled the survey, or many people can do it), each feedback relates to exactly one travel, with no optionality.

- ***planned_by*** (‘route’ **1:1...n** 'schedule’) - shows scheduled arrivals and departures for specific route (schedule differs for different type of a day: weekday, holidays etc.), each route at given day type has one or many schedules (e.g. different departure times) between two stops, each schedule relates to exactly one route, without optionality.

Relational Database Schema

Bus (RegistrationNumber, Type)

Travel (TravelID, IssuedDate, RegistrationNumber REF Bus, RouteNumber REF Route, ScheduleID REF Schedule)

Feedback (FeedbackID, SatisfactionLevel, TravelID REF Travel)

Validation (TravelID REF Travel, TicketID REF Ticket)

Ticket (TicketID, Name, Surname, Email)

Route (RouteNumber, RouteName, StartStop, EndStop)

Schedule (ScheduleID, ArrivalTime, DepartureTime, DayType, RouteNumber REF Route)

Additional EXCEL files

1. Service_Office EXCEL

Sheet 1 (Includes information about all bus service stations assigned to operating regions of the country, each line contain data about single bus service, line 1 is a row with headers)

Column A - Bus Service Identification Number (numeric, 0 decimal precision), indicate service that is responsible for given bus, its maintenance and travelled routes

Column B – Bus Service's Name (text)

Column C - Adress (text), includes information about street and house number,

Column D - Postal code (text),

Column E - City (text),

Column F – Region (text), single voivodship of Poland, indicate for what region given bus service is responsible,

Column G – Country (text), currently only Polish services,

Column H – Bus Slots (numeric, positive integer), number of parking spaces inside station.

Note: If the data about single station is updated e.g. address has changed or more bus slots are added, only specific column value of single row is updated. <u>One row always represents only one service station.</u>

Sheet 2 (Information about buses in all services offices, each line contain data about **<u>single</u>** bus, line 1 is a row with headers):

Column A – Bus Service Identification Number (numeric, 0 decimal precision), indicate service that is responsible for given bus, its maintenance and travelled routes,

Column B – Bus Registration Number (text), unique string consisting of 2-8 characters,

Column C – VIN (text), unique string consisting of 17 characters,

Column D – Brand (text),

Column E – Type (text), one of following options: *minibus, standard, low-floor,*

Column F – Production Year (numeric), format YYYY with 4 digits,

Column G – Seats (numeric, positive integer) number of sitting places,

Column H – Standing Places (numeric, positive integer) number of possible standing places,

Column I – Wheelchair (numeric, only 2 possible options: 0-false, 1-true), indicate a place for disabled person

Column J – Air Conditioning (numeric, only 2 possible options: 0-false, 1-true),

Column K – Feedback Monitor (numeric, only 2 possible options: 0-false, 1-true).

Note: If the data about single bus is updated e.g. air conditioning system is installed, only specific column value of one row is updated. The addition of new rows occurs only in case of adding new bus. One row always represents one vehicle.

2. Bus_Stops EXCEL

Sheet 1 (Information about all operating bus stops, each line contain data about **single** bus stop, line 1 is a row with headers):

Column A – Identification number of a bus stop (numeric, 0 decimal precision),

Column B – Bus Stop Name (text), related to its location – including city or airport name etc.

Column C – Longitude (double, 4 decimal places), information about geographical location of the bus stop,

Column D – Latitude (double, 4 decimal places), information about geographical location of the bus stop,

Column E – Sitting Place (numeric, only 2 possible options: 0-false, 1-true) indicate if there is a sitting place on a stop,

Column F – Airport (numeric, only 2 possible options: 0-false, 1-true) used to indicate that bus stop is close to airport,

Column G – Bus Shelter (numeric, only 2 possible options: 0-false, 1-true).

Note: If the data about single bus stop is updated e.g. bus shelter is added or bus stop is moved to other location, only specific column values of **one row** will be updated. The addition of new rows occurs only in case of adding new bus stops. One row always represents one bus stop.

Assumptions/Limitations: It is assumed that the records are kept from 2018 so it is the earliest year that can appear in database, additionally records will be kept in the database for a period of not less than 5 years (especially information about all bus travel and their validations). Some EXCEL values will certainly be updated during this time.

Scalability: Database needs to be designed to handle a potentially large volume of data as the number of routes travelled and validated tickets constantly grows.

3. Scenarios of analytical problems

Analytical problem: Why are there such differences in the number of passengers?

1. Compare the average number of passengers per route on weekdays and weekends in current and previous month.
2. On which days are buses most crowded: weekdays, weekends or holidays?
3. Identify routes with a 10% increase in average passenger count from the current month to the previous month.
4. Identify routes with the lowest bus occupancy (<15% of total bus capacity).
5. Compare the number of travelled routes from different bus offices in the current month to those in the previous month.
6. From which region did the bus office operate the most bus travel?
7. Compare the average number of passengers depending on the type of bus in current and previous month.
8. How bus stop's feature: "close to airport" have an impact on bus occupancy?
9. Is number of passengers related to distance between start stop and end stop?

Analytical problem: Why was there an increase and decrease in satisfaction in the surveys?

1. Compare the average satisfaction level for all routes in the current month to those in the previous month.
2. Whether passenger satisfaction level decreases when buses exceed a certain occupancy threshold? (<30%, 30-75%, >75%)
3. Do certain routes are associated with more frequent passenger feedback?
4. Are passengers more likely to complete surveys when travelling on a bus with a feedback monitor installed?
5. Which route has the lowest and highest average satisfaction level in the survey?
6. Do additional amenities such as air conditioning or wheelchair access affect satisfaction levels?
7. Are old buses (Production Year < 2010) receiving lower ratings?
8. Compare bus delays during weekdays, weekends and holidays.
9. How bus delays influence satisfaction levels?

4. Data needed for analytical problems

Analytical problem: Why are there such differences in the number of passengers?

1. Compare the average number of passengers per route on weekdays and weekends in current and previous month.
   - ***average number of passengers*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - ***all routes*** - *RoutesTraveller*, table *Routes,* column *RouteNumber*
   - ***type of a day (weekday or weekends)*** - *RoutesTraveller*, table *Travel,* column *IssuedDate*, information must be collected from some publicly available calendar, e.g. Google calendar
   - ***current and previous month*** - *RoutesTraveller*, table *Travel,* column *IssuedDate*

2. On which days are buses most crowded weekdays, weekends or holidays?
   - ***number of passengers (crowding)*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - ***type of a day (weekday, weekends, holidays)*** - *RoutesTraveller,* table *Schedule,* column *DayType*

3. Identify routes with a 10% increase in average passenger count from the current month to the previous month.
   - ***all routes*** - *RoutesTraveller*, table *Routes,* column *RouteNumber*
   - ***average number of passengers*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - ***current and previous month*** - *RoutesTraveller*, table *Travel,* column *IssuedDate*

4. Identify routes with the lowest bus occupancy (<15% of total bus capacity).
   - ***all routes*** - *RoutesTraveller*, table *Routes,* column *RouteNumber*
   - ***bus occupancy*** – calculated as ratio *number of passengers/total bus capacity*
   - ***number of passengers*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - ***total bus capacity*** – *Service_Office EXCEL, sheet 2* - sum of (Column *G – Seats* (numeric, positive integer) number of sitting places + Column *H – StandingPlaces* (numeric, positive integer) number of possible standing places)

5. Compare the number of travels from different bus offices in the current month to those in the previous month.
   - ***number of travels*** - *RoutesTraveller*, table *Travels,* column *TravelID,* check for entity existence
   - ***different bus offices*** - *Service_Office EXCEL, sheet 1,* Column *B – Bus Service's Name*
   - ***current and previous month*** - *RoutesTraveller*, table *Travel,* column *IssuedDate*

6. From which region did the bus office operate the most bus travels?
   - ***region*** - *Service_Office EXCEL, sheet 1,* Column *F – Region*,
   - ***bus travels*** - *RoutesTraveller*, table *Travels,* column *TravelID,* check for entity existence

7. Compare the average number of passengers depending on the type of bus in current and previous month.
   - ***average number of passengers*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - ***type of bus*** - *Service_Office EXCEL, sheet 2,* Column *E – Type*
   - ***current and previous month*** - *RoutesTraveller*, table *Travel,* column *IssuedDate*

8. How bus stop's feature: "close to airport" have an impact on bus occupancy?
   - close to airport feature - *Bus_Stops EXCEL, sheet 1,* Column *F – Airport* (compare cases when it is equal to 1-true and 0-false)
   - ***bus occupancy*** – calculated as ratio *number of passengers/total bus capacity*
   - ***number of passengers*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - *total bus capacity* – *Service_Office EXCEL, sheet 2* - sum of (Column *G – Seats* (numeric, positive integer) number of sitting places + Column *H – StandingPlaces* (numeric, positive integer) number of possible standing places

9. Is number of passengers related to distance between start stop and end stop of a route?
   - ***stops locations*** – geographical location is obtained from *Bus_Stops EXCEL, sheet 1* by combining both Column *B* and *C* (*Latitude* and *Longitude)*,
   - ***number of passengers*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - ***distances between bus stops*** - there's no such information in the given data sources. We propose obtaining such information:
     - using tool like *Google Maps* to calculate distances between stops on each route,
     - store distances for each route in *sheet 2* of *Bus_Stops EXCEL,* Columns: *StartStop and EndStop (*identification numbers of a bus stops from *sheet 1)* and *DistanceBetween*

   After such improvement we can use aggregate functions on calculated distance.

Analytical problem: Why was there an increase and decrease in satisfaction in the surveys?

1. Compare the average satisfaction level for all routes in the current month to those in the previous month.
   - ***average satisfaction level*** - *RoutesTraveller*, table *Feedback,* column *SatisfactionLevel*
   - ***all routes*** - *RoutesTraveller*, table *Routes,* column *RouteNumber*
   - ***current and previous month*** - *RoutesTraveller*, table *Travel,* column *IssuedDate*

2. Whether passenger satisfaction level decreases when buses exceed a certain occupancy threshold? (<30%, 30-75%, >75%)
   - ***satisfaction level*** - *RoutesTraveller*, table *Feedback,* column *SatisfactionLevel*, calculated for each threshold
   - ***bus occupancy*** – calculated as ratio *number of passengers/total bus capacity*
   - ***number of passengers*** – calculated from *RoutesTraveller,* table *Validation,* check for entity existence
   - *total bus capacity* – *Service_Office EXCEL, sheet 2* - sum of (Column *G – Seats* (numeric, positive integer) number of sitting places + Column *H – StandingPlaces* (numeric, positive integer) number of possible standing places

3. Do certain routes are associated with more frequent passenger feedback?
   - ***all routes*** - *RoutesTraveller*, table *Routes,* column *RouteNumber*
   - ***number of completed surveys –*** *RoutesTraveller,* count table *Feedback,* check for entity existence

4. Are passengers more likely to complete surveys when travelling on a bus with a feedback monitor installed?
   - ***number of completed surveys –*** *RoutesTraveller,* count table *Feedback,* check for entity existence
   - ***feedback monitor -*** *Service_Office EXCEL, sheet 2,* Column *K – FeedbackMonitor* (compare cases when it is equal to 1-true and 0-false)

5. Which route has the lowest and highest average satisfaction level in the survey?
   - ***all routes*** - *RoutesTraveller*, table *Routes,* column *RouteNumber*
   - ***average satisfaction level*** - *RoutesTraveller*, table *Feedback,* column *SatisfactionLevel*

6. Do additional amenities such as air conditioning or wheelchair access affect satisfaction levels?
   - ***additional amenities -*** *Service_Office EXCEL, sheet 2* **-** (Column *I – Wheelchair*) OR (Column *J – Air conditioning*) - compare cases with 1-true and 0-false values
   - ***average satisfaction level –*** calculated from *RoutesTraveller*, table *Feedback,* column *SatisfactionLevel*

7. Are old buses (Production Year < 2010) receiving lower ratings?
   - ***production year of a bus -*** *Service_Office EXCEL, sheet 2,* Column *F – Production Year*
   - ***satisfaction ratings*** - *RoutesTraveller*, table *Feedback,* column *SatisfactionLevel*

8. Compare bus delays during weekdays, weekends and holidays.
   - ***delays on stops*** – there is no such information.
   - ***type of a day (weekday, weekends, holidays)*** - *RoutesTraveller,* table *Schedule,* column *DayType*

9. How bus delays influence satisfaction levels?
   - ***is bus delayed –*** there is no such information.
   - ***average satisfaction level –*** calculated from *RoutesTraveller*, table *Feedback,* column *SatisfactionLevel*

To build proper BI system that solve all the analytical problems above, we suggest additional activities to be introduced. Bus operators should use GPS tracking systems that will be installed on buses. These systems should provide real-time data on the precise location and movement of each bus along its route. That enables to track arrival and departure times at stops – for passengers it might be shown through digital displays or mobile applications – and identify any delays according to the scheduled timetable. The data about delays in seconds will be automatically stored in the separate live-location system as to track delays on bus stops, what will be useful for further analysis. Sample structure of an excel sheet with an information about bus delays will contain:

Column A - TravelID (identify number, identifying one travelled route),

Column B - RouteNumber (identify number, by that we can find what is the scheduled time at bus stop)

Column C - BusStop(number)

Column D - BusArrival(datetime, data obtained from GPS system)

Column E - BusDeparture(datetime, data obtained from GPS system)

With that information in separate system, we can easily compare times and calculate delays. We can also put such information into “RouteTraveller” system or other EXCEL sheet.

**By orange color we indicate the queries that cannot be answered based only on data from RoutesTraveller system and Excel files (Service_Office & Bus_Stops).*