DATA MARKET
AUSTRIA

www.datamarket.at

Mobility
Pilot-specific Requirements

Deliverable number	*D8.1*
Dissemination level	*Public*
Delivery date	*03.11.2017*
Status	*Version 1.0*
Author(s)	*Martin Kaltenböck (SWC), Bernhard Niedermayer (Catalysts),*

The Data Market Austria Project has received funding from the programme "ICT of the Future" of the Austrian Research Promotion Agency (FFG) and the Austrian Ministry for Transport, Innovation and Technology (Project 855404)

Executive Summary

This document contains the requirements specification of the Data Market Austria (DMA) WP8 pilot that is about the topic of specification and implementation of a taxi demand heat map on the basis of mobility data (moving data of crowds), weather data, and additional open data (such as traffic situations, jams, or events).

Involved DMA partners in this pilot are: SWC, Catalysts, T-Mobile (+ Taxi 40100), ZAMG, Johanneum Research, and Siemens.

This pilot is one of (at least) two planned use cases, whereby this one is led by T-Mobile and Catalysts, while the 2nd one (to be specified later in the project) will be led by Siemens and Catalysts.

The objective of the pilot at hand is to design and realise an application that supports taxi fleet managers to better plan and operate taxi fleets in respect to positioning of taxis, as well as in the course of ‘where should a taxi go after a successful finalised taxi ride’. Thereby, costs and resources can be saved and revenue as well as profits of a taxi fleet can be increased.

This document specifies the use case, the data, and existing IT systems, as well as organisational, business and legal issues and finally describes the technical implementation of the pilot taxi demand heat map (service).

Table of Contents

1. Introduction

1.1 Title of Use Case (scenario, pilot)

Taxi Demand Heatmap

1.2 Pilot Domain (incl. partner)

Domain: Mobility

Involved Partners: SWC, Catalysts, T-Mobile (+ Taxi 40100), ZAMG, Johanneum Research, and Siemens

1.3 Problem Definition

Taxis are an integral part of the transport systems world-wide. Mainly in cities, taxis are widely-used and the competition between taxi fleet companies and/or individual drivers is high. The margin (profit) per ride is relatively small and thereby the level of utilization is one crucial factor for competitive advantage and business success in this industry.

Individual taxi drivers develop an experience on the optimal strategy after finishing a ride. Options are to stay and wait, to drive to the nearest taxi stand, or to drive to the nearest "hotspot", etc. Such hotspots are often well-known in the form of implicit knowledge of taxi drivers, but are only rarely-used by the coordinator of taxi fleets to optimise placement of taxis and/or to guide a taxi driver to the next and nearby hotspot after a managed ride.

Thus, operators of a taxi fleet (taxi fleet companies) have a strong demand to optimise the positioning (and guidance) of taxis of the fleet around the region the company is operating in (e.g. City of Graz, City of Vienna).

This can be supported by (i) predicting the demand for taxi rides for the next 15 minutes in a given grid, (ii) thereby identifying relevant hotspots in a given grid (along different event types, people crowds, moving data of people (crowds), etc), and (iii) by giving the fleet operator and taxi driver information about the best way to serve such demand (route, position, etc) and in consequence reach the highest possible taxi fleet utilization.

1.4 Purpose & Objectives

The list of objectives and possible achievements of this pilot are as follows:

- Raising the use / utilization of a taxi (cab) fleet - optimising the fleet utilization
    - Measured by: number of (overall) managed taxi rides
    - Measured by: Total amount of idle time (time taxis of the fleet are unused / have no customers)
- Improved scheduling of entire taxi fleets
- Improved resource planning

- Growth of the company employing this service (software provider for taxi fleets and/or taxi fleet operator / company)
    - Optimised prediction where to position taxis / where taxis are needed, therefore more taxis in the fleet possible, potentially prices can be improved, more revenue can be generated.
    - Also to be taken into account: less taxis can carry the same amount of passengers.
- Outlook: in the course of 'taxi sharing' a lower cost per taxi user can be achieved; this implies also more taxi use and users over time
- Vision: with the emerging technology of autonomous cars, the driver's implicit knowledge about hotspots or strategies for profit optimization in general will fall away. At that time, computer systems shall be able to take over and control the taxis.

1.5 Scenario

In practice, several questions arise that can be answered automatically, using the taxi demand heatmap. These questions include (non comprehensive list):

- What is the optimal strategy for a taxi (as part of a fleet) that has just ended its current ride (stay & wait versus go to another place)?
- Where should taxis be placed when starting the working shift?
- Should taxis be replaced as of changing conditions in regard to weather, arising hotspots, traffic issues, or other events?

2 Data of the Pilot & IT Prerequisites

2.1 Primary Content / Data Involved

- Past taxi rides in the Vienna area (40100):
    - GPS track
    - distance
    - driving time
    - starting/end location
    - starting/end time
- Distribution of people over the area (T-Mobile):
    - number of persons per grid cell
- Weather conditions (ZAMG):
    - temperature
    - rainfall
- Train delays (public):
    - minutes of delay
- Events (public):
    - type of event
    - location
    - start/end time
    - expected number of visitors
- Traffic jams and traffic / road conditions as for instance construction places (public):
    - expected delay in minutes
- Time of day, day of week, season (public)

2.2 (Meta) Data Structures

The data listed above also provide respective metadata - as follows a list of metadata per relevant data:

- Distribution of people: grid, time,
- Weather conditions: grid, time
- Train delays: arrival station location, expected number of passengers
- Events: type of events (classification of events)
- Traffic jams and road conditions: grid, address, latitude / longitude, type of jam / condition (classification)
- Taxi fleets: grid, base address, time of operation, number of cars

2.3 IT Prerequisites

The following prerequisites come along with different data providers:

Taxi 40100

- As Taxi 40100 is not a consortium partner, this information could not be shared / is not really relevant for this document.

T-Mobile

- T-Mobile Austria hosts a Hortonworks Data Platform (HDP), to form a coherent, secure, governed, and managed *data platform* for ingesting, preparing, and provisioning of arbitrary data sources. A multi-layer staging concept transforms incoming raw data into use-case specific data, while a sophisticated access control scheme guarantees safe delivery and handling.

  - Hardware
    The scoping for the hardware order resulted in a set of requirements for the initial Hadoop cluster, which needs to store close to a Petabyte of data. While the order process is still ongoing and exact details are not available, the requirements asks for about 30 servers, with state-of-the-art specifications.

  - Network
    For the T-Mobile data platform, a specific network setup was chosen, which ensures strong security guarantees by dividing the access layer from the actual cluster network. Figure 1 shows an overview of how the various components of the platform are assigned to specific networks.

  - Software
    The initial version of data platform should follow the KISS principle, which means in a nutshell to keep things *simple* and *predictable*. As mentioned above, the Hadoop platform and its components are still in flux and often change quite extensively across subsequent releases – which usually occur multiple times a year. On the other hand, the core Hadoop ingredients, such as the file system HDFS and the resource manager YARN, are considered stable and are updated with more predictable

changes. Figure 2 shows an overview of the various components used in the design of the T-Mobile data platform.

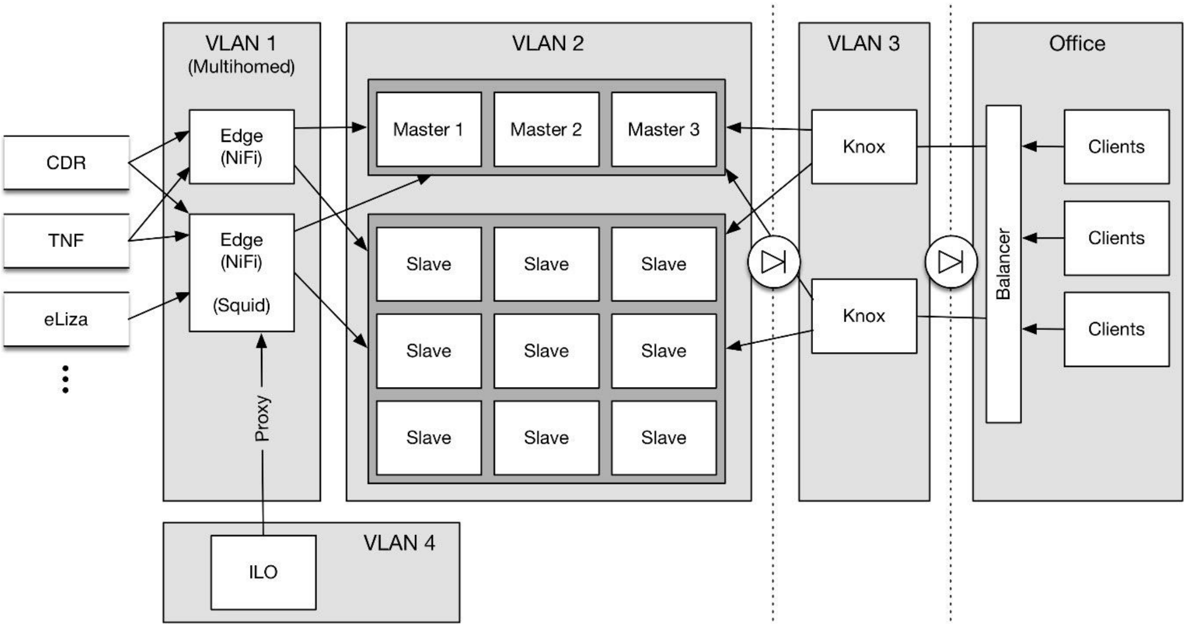


Figure 1: Network setup of components within T-Mobile data platform

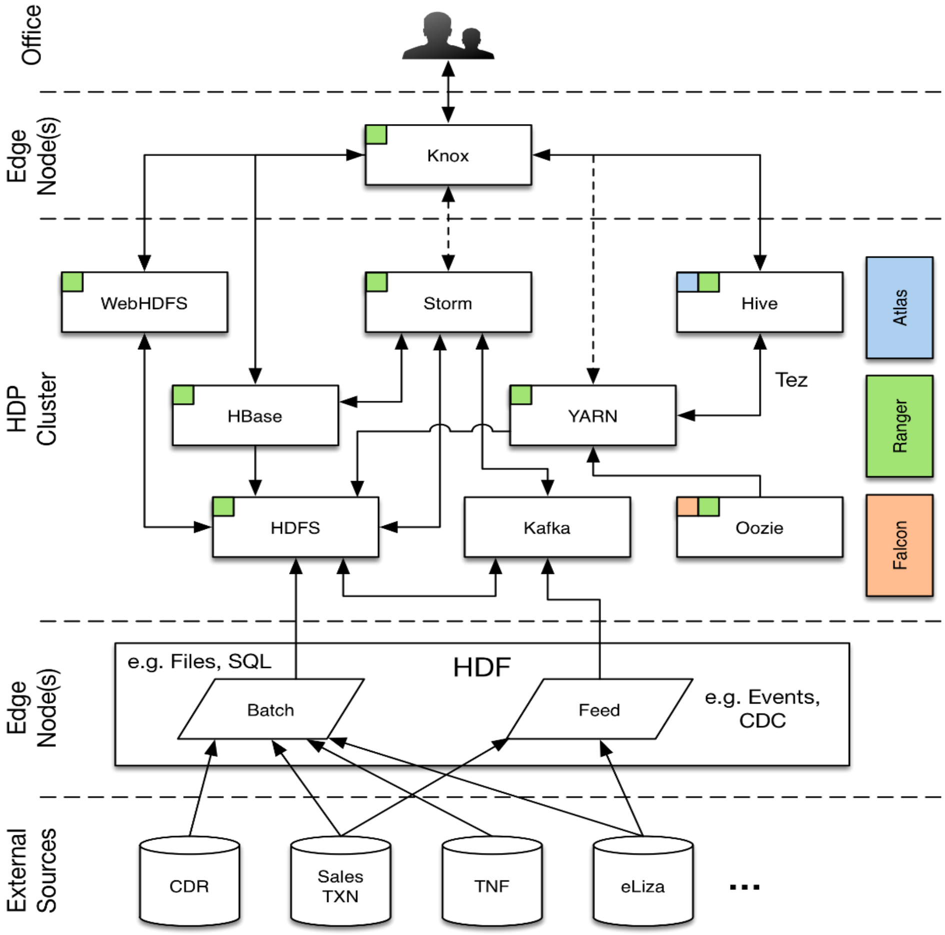


Figure 2: Overview of the T-Mobile Data Platform

ZAMG

- ZAMG uses an API management architecture, based on an API gateway and microservices to deliver data over the Internet (see Figure 3). The data are stored in relational databases, NoSQL databases, and file-based archives. For data in file-based archives, the GRIB and NetCDF formats are used. In order to access the data, open source software is used to provide RESTful Web Services. The services are implemented according to the microservice design principle and are made available to the user via the API management system. In the management layer, the API gateway provides a unified user interface and enables centralized implementation of analytics, caching, and authentication/authorization services.

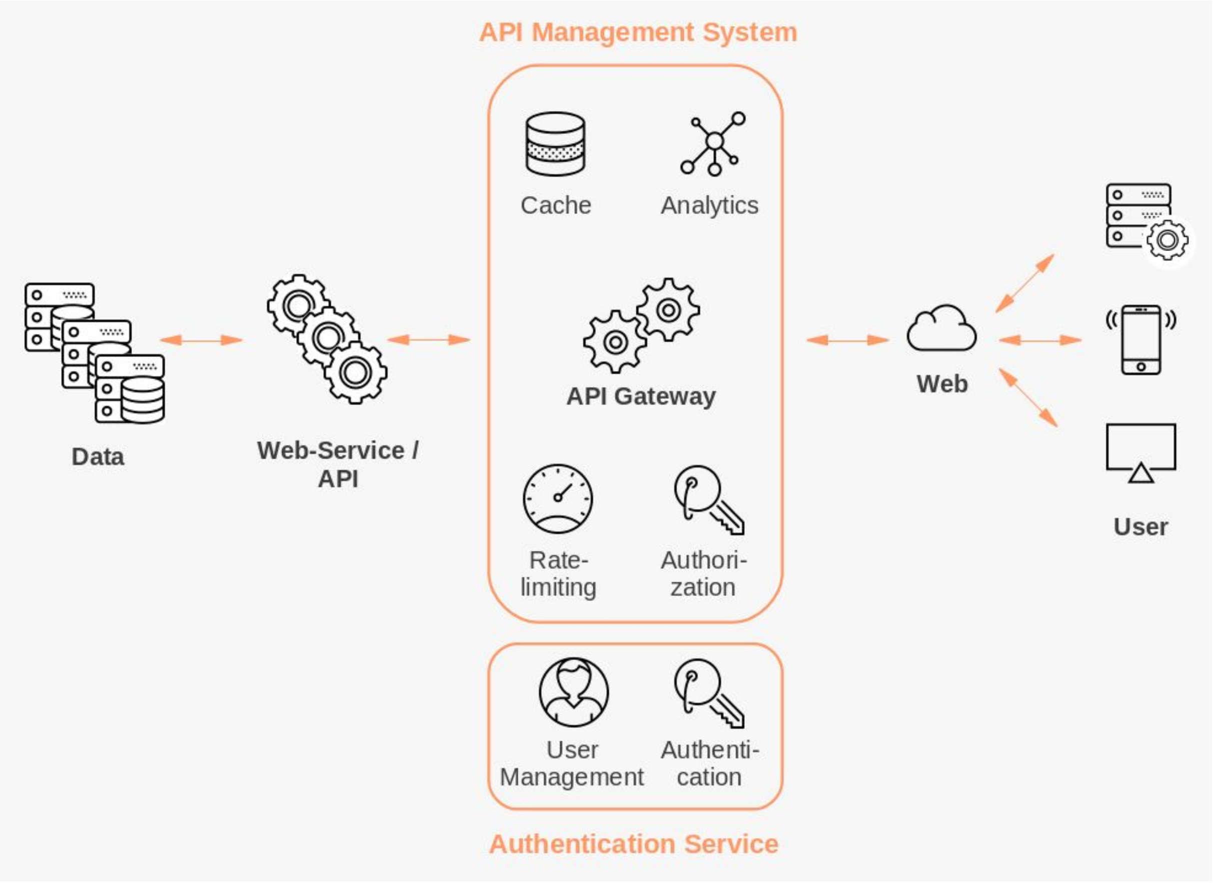


Figure 3: API Management Architecture

2.4 Existing IT-Systems

Taxi fleet operators facilitate fleet control systems with basic features, such as

- Register incoming orders for rides
- Track car positions
- Assign ordered rides to cars
- Register spontaneously picked up riders

Fleet control systems consist of a common backend and various clients for call-center, drivers, fleet control and optimization, etc. These clients can be implemented as desktop applications as well as mobile apps.

Furthermore, data will be provided mainly by T-Mobile (as a service) and ZAMG, based upon their existing IT systems.

3 Organisational & Legal Aspects of the Pilot

3.1 Role Models Involved

The following main stakeholders are involved in this use case:

- Taxi fleet operator
- Taxi driver
- Passenger

3.2 Impacts

This pilot can change the way, taxi fleets (and similar businesses) are maintaining their operation mainly by (improved) prediction of relevant events and thereby changing the way of placing taxis of the fleet in a pre-defined area / region. This can be a best-practise example for this domain and can raise revenues and taxi use.

3.3 Workflows

- Automatic instead of manual fleet control and optimization
- Driver is advised about best strategy to pick up the next ride
- Driver is advised about best strategy where to position a taxi (and also when changes occur)

3.4 IT Infrastructure

- Taxi demand heatmap as external service via DMA
- New Interfaces for data input / output

- Automatic scheduling and controlling algorithm

3.5 IPR / Legal Certainty

Several questions arise, concerning ownership of data products and services:

- Who will own the taxi demand heatmap?
  - This answer is still in clarification at the time of this document being created
- How will contributing data providers participate?
  - Data will be provided for the pilot in the course of the DMA project, but commercial data needs to be part of the service business model after the DMA project is finalised
- Which licensing model shall be applied?
  - Commercial including SLAs that needs to be specified
- Additional Criteria required?
  - Data needs to be 100% aggregated and anonymised
- The concrete business model of such a service needs to be developed, taking into account:
  - DMA infrastructure costs
  - DMA basic services costs
  - Costs for data and data services (ZAMG, T-Mobile, Taxi40100)
  - Costs for newly developed DMA services (if any)

3.6 Revenue & Cost Structures

The following list describes the parameters of the business model about the new Taxi Demand Heat Map Service:

- Revenue increases due to minimized idle times
- Potentially decreased operating costs due to automatic scheduling instead of manual
- Increased costs due to additional software components
- The concrete business model of such a service needs to be developed, taking into account:
  - DMA infrastructure costs
  - DMA basic services costs
  - Costs for data and data services (ZAMG, T-Mobile, Taxi 40100)
  - Costs for newly developed DMA services (if any)

4 Risk Analysis

The following risks (technical risks and business risks) could be identified in the course of the requirements specification phase of this pilot.

4.1 Technical Risks

- Input data not available as required (or incomplete)
- Computations must meet performance goals, with the quantity of input data can grow large
- Prediction of events and thereby positioning of taxis not adequate

4.2 Business Risks

- Bad recommendations made by automatic scheduling - intrinsic knowledge of taxi drivers outperforms the computed recommendations
- Increased costs for software / data are not amortized by increased revenues
- Legal issues of not sufficient anonymisation of data

5 Non-Functional Requirements

This section is structured (if information available) following the FURPS+ model[1]:

- **Functionality**: security, future extensibility, capability …
- **Usability**: aesthetics, documentation, consistency …
- **Reliability**: failure rate, recoverability, data consistency ….
- **Performance**: responsiveness, scalability, throughput …
- **Supportability**: testability, maintainability, adaptability …

Furthermore ‘Portability’ might be an issue that needs further investigation to ensure easy adoption of the pilot technology / application / source code for re-use / a generic service.

5.1 Functionality

Security

- The API providing the taxi demand heatmap is protected by the token-based DMA security service.

Extensibility

- The heatmap will be provided for the Vienna area.

[1] https://en.wikipedia.org/wiki/FURPS

- Prediction models for further cities can be trained given the necessary input data.
- Prediction models including additional data sources can be trained, but require an additional data analysis phase.

5.2 Usability

- The API will be documented using the Swagger format and UI
- The taxi demand heatmap service will implement the DMA Service API, such that it can be operated and managed by the DMA platform
- The API will follow the DMA style and naming conventions

5.3 Reliability

- The data quality depends on the input data and the quality of the trained model
- The service itself is stateless, i.e. recoverable without additional mechanisms

5.4 Performance

- Runtime performance has to be considered for the prediction step (not the model training step that is done offline).
- Update rate of 1/minute
- The considered geographical range is the Vienna area

5.5 Supportability

Support for customers of the heatmap service (software developers) will be given in the form of documentation following the OpenAPI format.

6 DMA Pilot-specific Technical & Data-related Specification

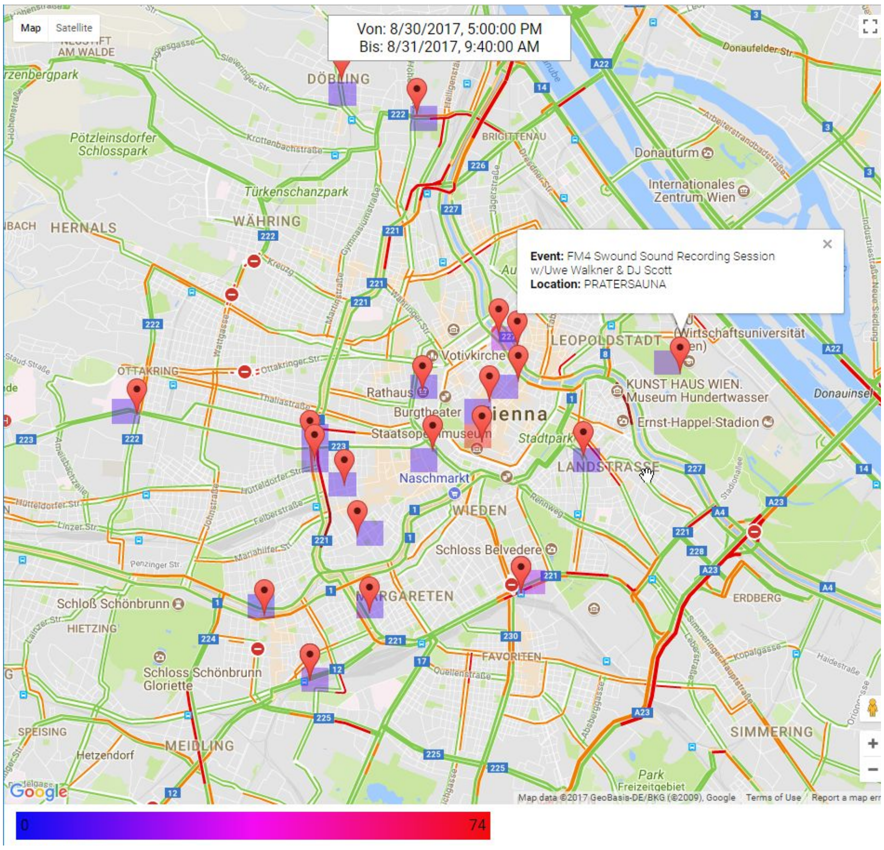
6.1 Technical Description of Solution

The taxi demand heatmap service will be introduced as a new component. Its purpose is to combine several data sources to predict the demand for taxi rides in the Vienna area. The computation of this demand will be based upon a model, which is trained in a preceding data analysis step.

The data analysis will be done using the Python stack (including libraries such as numpy, matplotlib, scikit learn, pandas, etc.). Geo-referencing is done using gdal.

According to this, the prediction component will be done using the Python stack as well. Selection of the exact web service technology is within the scope of the project.

The following prototypical visualization shows a first approach where event locations are assumed to have an increased demand for taxi rides.



Von: 8/30/2017, 5:00:00 PM
Bis: 8/31/2017, 9:40:00 AM

Event: FM4 Swound Sound Recording Session w/Uwe Walkner & DJ Scott
Location: PRATERSAUNA

6.2 Interdependencies of IT-Components

The taxi demand heatmap service relies on following static data acquired once in order to train a prediction model:

- Mobility data
- Taxi rides data
- Historic weather data
- Historic train delays
- Historic events

Dependencies on the following systems exist during production:

- Live mobility data
- Weather forecast
- Live train delays
- Live events

The service to be developed will integrate into DMA by implementing the DMA Service API.

The service itself offers a documented API to be integrated into taxi fleet control software.