Vorstudie Künstliche-intelligenz basierte Belegungsoptimierung im ÖPNV

- **Fahrbar20 // 19F1145A** -

Schlussbericht

Walldorf, November 2024

Projektleitung: Dr. Tim Frey | tim.frey@iunera.com

iunera Gmbh & Co. KG | Altrottstraße 31, 69190 Walldorf |

Inhaltsverzeichnis

1. Aufgabenstellung.....2

2. Voraussetzungen.....2

3. Planung und Ablauf.....3

4. Arbeitspakete.....4

4.1 Datenplattform.....4

4.2 Intelligente Linienzuordnung.....4

4.3 KI Vorhersagen.....4

4.4 Veröffentlichung und Konferenzteilnahme.....5

4.5: Projektfortführung.....5

4.6 Mobilitätsdatenverordnungs-konforme Aufbereitung der Vorhersagen.....5

5. Stand von Wissenschaft und Technik.....6

5. Zusammenarbeit.....8

6. Zielgegenüberstellung.....8

7. Mittelverwendung und Angemessenheit.....9

8. Innovative Verwertungsoffensive durch Blockchain basierte Lizenzierung.....11

8. Nutzen und Veröffentlichungen.....12

8.1 Linienzuordnung zu GPS Daten.....12

8.2 Weiterentwicklungen zu Apache Flink.....13

8.3 GTFS Service.....13

8.4 Datentypen und Schlüsselgenerierung.....14

8.5 Apache Druid Erweiterungen.....14

8.6. Daten Anreicherung.....15

8.7 Daten Streamifizierer.....15

8.8 Infrastruktur als Code.....15

8.9 Konferenzbeiträge.....16

9. Entwicklungen während der Durchführung des Vorhabens.....16

10. Fazit.....17

1. Aufgabenstellung

Analyse, Vorhersagen von Belegungen im ÖPNV sind unentbehrlich um Fahrzeuge intelligent und Co2 optimiert einzusetzen. Äußere Einwirkungen wie Benzinpreisänderungen, Ticketingaktionen oder Jahreszeiten machen jedoch zuverlässige traditionelle Belegungsvorhersagen unmöglich. Besonders im ländlichen Raum ist es schwierig Belegungsanalysen durchzuführen, da die ÖPNVs hier über geringere finanzielle Mittel verfügen. Es wird somit eine Softwareinnovation benötigt die besonders bei kleinen ÖPNVs anwendbar, anpassbar und erweiterbar ist.

Es fehlt an einer standardisierten offenen Softwareplattform um Sensordaten zu erfassen, aufzubereiten, zu analysieren und breitflächig Vorhersagen erstellen zu können. Dieses Projekt konkretisiert und entwickelt daher die zugehörige Dateninfrastruktur und Künstliche Intelligenz (KI)-Vorhersagemechanismen.

2. Voraussetzungen

ÖPNVs sind derzeit weitestgehend nicht ausreichend mit Personenzählgeräten, Datenplattformen und KI Wissen ausgerüstet. Weiterhin ist der ÖPNV stark fragmentiert. Jedoch ergibt sich die Herausforderung, dass ÖPNVs detaillierte Belegungsanalysen und Vorhersagen oftmals erst durchführen, wenn mehr Personenzählsysteme installiert sind. Die Installation von mehr Personenzählsystemen ist aber ohne erweiterte Geschäftsanwendungsfälle organisatorisch schwierig.

Datenprojekte wie Data Lakes, KI und ähnliche Projekte scheitern in großen ÖPNV Betreiber- und Gesellschaften, oftmals an nicht verfügbaren Fachkräften und an Open Source Leuchtturmprojekten für die Fachdomäne, die kopiert und erweitert werden können, um Eigenentwicklungen auf der grünen Wiese abzukürzen. Für kleine ÖPNVs ist es oft nicht im Aufwand umsetzbar, Projekte mit Belegungsdaten zu starten oder gar notwendige moderne Serverinfrastruktur einzurichten.

Vereinfacht gesagt ergibt sich ein Henne-Ei Problem Belegungsdaten vermehrt aufzunehmen und opensourcebasierte, übertragbare Datenplattformen zu erschaffen und damit die Verkehrswende datengetrieben sowie transparent zu steuern.

Zusätzlich erzeugen schnelle Wechsel von Verhaltensmustern, wie sie im ÖPNV vorkommen, Herausforderungen im KIs Forschungsbereich. Gesamtheitlich betrachtet zeigt die fachliche Ausgangslage KI, dass der Bedarf an adaptiven KI Lernmethoden besonders hoch ist, was somit die KI Entwicklung stärkt.

Datentechnisch ergeben sich zusätzliche Herausforderungen. Belegungsdaten im ÖPNV werden durch Belegungssensoren in Fahrzeugen generiert. Hierbei messen Sensoren die Einsteiger und Aussteiger aus einem Fahrzeug und diese müssen dann Fahrplandaten zugeordnet werden. Diese resultierenden Belegungsdaten werden dann durch Fahrgastzählungen validiert und dienen dem ÖPNV dann hochgerechnet für die Einnahmenaufteilung. Problematisch ist es, dass Fahrzeuge auf verschiedenen Linien eingesetzt werden und die Zuordnung von Fahrzeugen zu Linien dann im Regelfall manuell durch den ÖPNV durchgeführt wird. Für die Einnahmenaufteilung reicht normalerweise eine geringe Anzahl von Sensoren und gemessenen Fahrten aus. Für Vorhersagen sind aber

wesentlich mehr Datenpunkte notwendig, aber es ist derzeit unklar welche Rahmenbedingungen sich an Datenpunkte ergeben, damit Vorhersagen generell möglich sind.

3. Planung und Ablauf

Es wird eine Cloud agnostische Infrastruktur entwickelt, die Datenintegration, Datenströme, Stammdaten und Künstliche Intelligenz Mechanismen als Dienste aufnehmen kann. Hierbei ist das Ziel, eine Code basierte Architektur zu entwickeln, die erweiterbar und an verschiedene Szenarien anpassbar ist.

Da oftmals Sensordaten, ohne Nachbearbeitung, keinen Fahrtlinien zugeordnet sind erfolgt im nächsten Schritt, nach der Entwicklung der Infrastruktur und der Integration von Daten, ein Matching von Sensordaten zu Fahrplandaten. Dieser Machbarkeitsbeweis ermöglicht es konzeptionell auch (Smartphone) GPS Daten ÖPNV Fahrten zuzuordnen.

Sofern keine klare Zuordnungen von Zähldaten zu Linien verfügbar sind, ist es geplant die Daten aus Zählsystemen mit Fahrplandaten zu verschneiden und somit klare und eindeutige Fahrzuordnungen möglich zu machen. Um dieses Ziel zu erreichen werden wir zuerst erweiterte, statistische Zuordnungsmethoden und falls notwendig starke KI Mechanismen einsetzen.

Die bereinigten Daten werden danach als Datenströme abgebildet und in einem speziell entwickelten Datenmodel verarbeitet.

Nach der Datenbereinigung werden die Daten mit „engineerten“ Meta-Merkmalen erweitert (z.B. Wochentage, Ereignisse, Relationen zu anderen Linien usw.).

Diese vereinheitlichten erweiterten Daten bilden im weiteren Vorgehen die Grundlage und statistische und Künstliche Intelligenz basierte Vorhersagen zu entwickeln.

Fortfolgend werden Vorhersagen durchgeführt und die Ergebnisse werden validiert und bewertet. Hierbei wird geprüft welche Merkmale ausschlaggebend sind, um adaptive Vorhersagen zu bestimmen. Die Merkmale werden interaktiv konkretisiert und erweitert.

Die notwendige technische Umsetzung basiert auf Apache Kafka, Spring Java Microservices, Apache Flink und unter der Nutzung von standardisierten Fahrplandaten (GTFS).

Sämtliche relevante Quelltexte werden veröffentlicht werden um Publikationen auf Fachkonferenzen zu komplementieren.

Es wird zusätzlich angestrebt auch die veredelten Daten bei einer Fortführung des Projektes, sofern eine Finanzierung sichergestellt ist, kontinuierlich zu veröffentlichen.

Hierzu soll im Projekt evaluiert werden wie das Projekt am besten fortgeführt werden kann und das Projekt soll einen Einblick über die Kapitalanforderungen und benötigte Planungssicherheit zur Fortführung geben. Gegebenenfalls soll ein Projektkonsortium gebildet werden um die Projektergebnisse zu verwerten und anzuwenden.

4. Arbeitspakete

Im Folgenden finden sich die geplanten Arbeitspakete und deren letztendliche Resultate beim Projektausgang.

4.1 Datenplattform

Es wird auf der Basis von Open Source Technologie eine Infrastruktur aufgesetzt. Hierbei ist das Ziel, dass die Konfiguration der Open Source Technologie revisioniert wird. Damit wird gewährleistet, dass die Infrastruktur selbst als Open Source Projekt geteilt, kopiert und weiterentwickelt werden kann.

Resultat:
Die Ergebnisse des Arbeitspaketes sind online in unseren Github Repositories[1] verfügbar und können weiterverwendet werden.

4.2 Intelligente Linienzuordnung

Wir entwickeln Methoden um Fahrpläne (in GTFS Format) automatisiert Belegungs-Sensordaten zuzuordnen. Dies ist herausfordernd, weil Verspätungen oder auch Linienwechsel von Fahrplänen hierzu erkannt werden müssen. Unser Ziel ist es, dass eine Zuordnung auch nahe der Echtzeit geschehen kann, wenn die Daten in Echtzeit geliefert werden. Zugleich kommunizieren wir die Zwischenergebnisse zu ÖPNVs und bewegen diese dazu, ihre Belegungsdaten öffentlich zur Verfügung zu stellen.

Resultat:
Die Ergebnisse des Arbeitspaketes sind online im Github Repository verfügbar und können weiterverwendet werden. Insgesamt sind die Ergebnisse ein Microservice für GTFS Daten und Streamprozessoren für Apache Flink. Wir testen unsere Datenbereinigung und Linienzuordung mit den Daten von einem ÖPNV und können positive Datenanreicherungen durchführen.
Durch die Anforderungen auch Echtzeitdaten verarbeiten zu können existieren zudem noch verschiedene Flink Erweiterungen und logische Datentypen für den ÖPNV erstellt. Auch diese sind als Open Source Projekte verfügbar.

4.3 KI Vorhersagen

Wir bearbeiten Daten mit Data Science Methoden und testen Apache Druid als Grundlage für Vorhersagen. Zugleich erstellen wir Jupyter Notebooks und engineeren Features darin.

Resultat:
Wesentliche Ergebnisse des Arbeitspaketes sind online in unserem Github Repository verfügbar und können weiterverwendet werden.
Insgesamt konnten wir gute Ergebnisse mit Quantil-basierter Regression mit den Statistik Modulen in Druid erzielen. Für Machine Learning basierte Vorhersagen zeigten sich aufgrund der dünnen Datenlage Random-Forest-basierte Modelle erfolgreich.
Essentielle Lehren aus den verschiedenen (nicht)erfolgreichen Versuchen haben wir auf Fachkonferenzen vorgestellt.

---

[1] https://github.com/iunera
https://iunera.com/

4.4 Veröffentlichung und Konferenzteilnahme

Hier werden die Ergebnisse auf Veröffentlichungen vorbereitet. Dies bedeutet, dass Kommentare, Beschreibungen und Qualitätssicherung für das Veröffentlichen von Quelltext und anderen Softwareartefakten durchgeführt wird.
Zusätzlich bewerben wir uns als Sprecher bei Konferenzen, die zu dem Themenfeld passen und erstellen Präsentationen und tragen diese vor.
Dies ist ein kontinuierlicher Prozess und das Ziel ist die Ergebnisse bei möglichst vielen Konferenzen vorzustellen.

Resultat:
Wesentliche Ergebnisse wurden als Open Source komplett transparent veröffentlicht. Weiterhin wurde an mehr Konferenzen teilgenommen als geplant. Vom Feedback aus KI Fachkonferenzen erfuhren wir reges Interesse an den entwickelten Methoden. Bei einer ÖPNV Fachkonferenz war das Interesse zwar auch gegeben, aber zugleich kamen auch viele Argumente warum man das alles nicht direkt so umsetzen sollte.

4.5: Projektfortführung

Wir versuchen ein Projektkonsortium zusammen zu stellen um das Projekt auf die nächste Stufe zu heben. Hierbei sollen durch das Konsortium die Ergebnisse der ersten Projektstufe die Nachhaltigkeit garantieren und Skaleneffekte auf andere ÖPNVs bewirkt werden.

Resultat:
Durch eine Blockchainbasierte Open Source Lizenz wird das Projekt nachhaltig gestaltet.[2]
Uns war es vor allem wichtig, dass Weiterentwicklungen unter der Auflage einer fairen Vergütung aller Entwickler stattfinden können, da die meisten Open Source Projekte daran scheitern, dass die Entwickler nicht oder bei weitem nicht ausreichend für ihre Arbeit vergütet werden. Weiterhin ermöglicht die innovative Blockchainbasierte Lizenz neue Kollaborationsmodelle zwischen verschiedenen ÖPNVs.

4.6 Mobilitätsdatenverordnungs-konforme Aufbereitung der Vorhersagen

Ein oder mehrere ÖPNVs erteilen die Datenfreigabe von Veröffentlichungen.
Technisch werden Vorhersagen Mobilitätsdatenverordnung-konform aufbereitet.

Resultat:
Organisatorisch standen wir vor der Herausforderung, ob es uns wichtiger ist den Quelltext aufzubereiten um diesen zu veröffentlichen oder direkt einen Upload in die Mobilithek zu programmieren, da nicht beides ressourcentechnisch möglich war. Wir entschieden uns dafür die Quelltexte zu veröffentlichen, da uns dies nachhaltiger erschien, da dies es erlaubt unsere Arbeit weiterzuentwickeln. Somit wichen wir hier etwas vom ursprünglichen Plan ab.

Vorbereitend für die Interaktion für mit Mobilithekdaten haben wir Schnittstellen definiert. Diese können als Grundlage verwendet werden um definierte Datenexporte für die Mobilithek zu erzeugen. Hierbei wurde ein Vorhersage Schnittstelle erschaffen, die wir genutzt haben um

[2] https://license-token.com/

Vorhersagen für Echtzeitauskunftssysteme zu erzeugen. Diese Arbeit ist als Source Code veröffentlicht und kann als Grundlage für Mobilitheksexporte verwendet werden.

5. Stand von Wissenschaft und Technik

Als verwandte Arbeiten können Vorhersagen aufgrund von Verbindungsauskünften oder Ticketverkäufen gesehen werden.
Es gibt zudem Schnittmengen zu unseren eigenen bisherigen Arbeiten und zu dem Themenfeld Data Stream Mining (wie z.B. Vorhersagen von Kryptowährungskursen).

Es gibt viele insbesondere wissenschaftlich wichtige Arbeiten, die wir gerne an dieser Stelle darstellen möchten, was jedoch die „ÖPNV Zielrichtung“ sprengen würde[3]. Daher fokussieren wir im Folgenden die einschlägigsten angewandten Arbeiten:

Vorhersagen mit Verbindungsanfragen, wie beispielsweise durch mFund– ProTrain

Es gibt im Fernverkehr und in Berlin Versuche Vorhersagen aufgrund von Verbindungsanfragen oder Ticketverläufen durchzuführen. Mathematisch ist dies eine einfache Statistik, die aufgrund von Verbindungsanfragen und Ticketverkäufen hochmultipliziert wird.
Unser Ansatz unterscheidet sich neben dem „Open Source Konzept“ hiervon, da es bei uns insbesondere um den Nahverkehr geht. Zum einen werden hier weitaus weniger relevante Verbindungsanfragen gestellt und zum anderen ist unser spezielles Ziel dabei Benutzungsmusteränderungen durch KI zu berücksichtigen. Weiterhin sind im Nahverkehr Umsteigevorgänge stärker zu beachten und die Vorhersagen müssen daher aufgrund von Personenzählgeräten bestimmt werden.

Vorhersagen aufgrund von Mobilfunkdaten, wie beispielsweise mFund - Mobile Data Fusion

Es gibt Versuche Mobilitätsvorhersagen aufgrund von Mobilitätsdaten vorzunehmen. Dies ist aber datenschutzrechtlich problematisch, wenn die Daten detailliert genug sind und bei unscharfen Daten ergeben sich Genauigkeitsprobleme. Zuletzt sind Mobilfunkdaten aufgrund von der Datenquelle potenziell nicht in Echtzeit verfügbar, wie dies bei Personenzählern in Zukunft der Fall sein wird. Dennoch können sich hier in Zukunft große Synergieeffekte ergeben, wenn genügend Zähldaten vorhanden sind und dann durch Mobilfunkdaten kalibriert

[3]Die Folgenden Publikationen halten wir für unser Projekt als besonders relevant/verwandt:
-Hermann Blümel, Mobilitätsdienstleister ohne Kunden. Kundenorientierung im öffentlichen Nahverkehr Berlin, 2004, WZB
-Barbara Lenz, *Mobilität nach Corona.* Mobilität nach Corona, Wie Corona unser Verkehrsverhalten verändert, 2020, DLR-Online.
-Lenz, Barbara (2020) *Richtung Zukunft Mobilitätskonzepte für die Stadt von Morgen.* Berlin Urban Tech Summit 2020
-Claudia Nobis, *Verkehrsmittelwahl bei Freizeitwegen: Analyse auf Basis von Modalgruppen.* In: Nahverkehrs-Tage 2015
-Tirachini et al., Crowding in public transport systems: effects on users, operation and implications for the estimation of demand Transport 2015
-Kim, Jisun, Revell, Kirsten and Preston, Jonathan, Application of cognitive work analysis to explore passenger behaviour change through provision of information to help relieve train overcrowding. In *Advances in Human Factors of Transportation: Proceedings of the AHFE International Conference on Human Factors in Transportation, 2019*
-Sanmay Shelat, Oded Cats, Sander van Cranenburgh, Traveller behaviour in public transport in the early stages of the COVID-19 pandemic in the Netherlands, Transportation Research Part A: Policy and Practice, 2022

werden. Weiterhin kann die Linienzuordung des Fahrprojektes potenziell dazu dienen Mobilfunkdaten Linien zuzuordnen.

mFund -Prognose und Visualisierung von Belegungsdaten auf Basis KI-gestützter Analyse – NV-ProVi

Im Gegensatz zum direkten Daten- und KI-Fokus von Fahrbar20 ist das Ziel von NV-ProVi die Visualisierung eines Verbandes. Hierbei soll ein lernendes System entwickelt werden, welches ÖPNV-Nutzern zu jeder Zeit eine möglichst genaue Prognose über Belegung und Pünktlichkeit der Busse und Bahnen, auf ihrer geplanten Route, zur Verfügung stellt. Diese Informationen sollen in ansprechender, intuitiver Form über eine Live-Map des Rhein-Main-Verkehrsverbundes visualisiert werden. Such- und Filterfunktionen sowie Routenkriterien sollen Fahrgästen ermöglichen, ihre individuelle Route möglichst nah an den eigenen Anforderungen zu planen. Analysen sollen dabei durch einen Data Lake geschehen.
Das NV-ProVi Projekt ist besonders spannend, da unsere Datenbereinigungsmechanismen, Aufbereitung, Zuordnung und Adaptive Vorhersage KI ergänzend sind.
Beispielsweise unterscheidet sich Fahrbar20 von NV-ProVi insbesondere dadurch, dass wir kleine ÖPNVs durch einfache Softwareelemente und Cloud Software As a Service enablen werden, wohingegen NV-ProVi auf einen großen Data Lake setzt.

mFund -Automatische Fahrgastzählung im ÖPNV mit Kameras und intelligenter Softwares

Durch videokamerabasierte Systeme werden Belegungen festgestellt und es wird das Problem adressiert, dass zu wenige Zählsysteme vorhanden sind. Potentiell lassen sich hier Projektergebnisse mit dem Fahrbar Projekt kombinieren indem die Belegungen dann mit unseren Prognosen und Linienzuordnungen kombiniert werden.

mFund -KI gestützte Schadstoffprognose zur intelligenten Verkehrssteuerung – KISP

Die in diesem Projekt erstellten Schadstoffprognosen können durch unsere Projektergebnisse potenziell zukünftig gemeinsam mit den von uns angestrebten Belegungsvorhersagen verknüpft werden. Damit kann die intelligente Verkehrssteuerung auf den ÖPNV angewendet werden.

Patentanmeldung EP22190257.0 um Fahrten aus verschiedenen Systemen abzugleichen

Die Patentanmeldung offenbart ein Verfahren, dass es ermöglicht GPS Daten trivial Fahrplandaten zuzuordnen. Dies wird im Projekt verwendet werden um Daten zu homogenisieren.

5. Zusammenarbeit

Die Zusammenarbeit im Projekt besteht darin, dass iunera von ÖPNVs Daten für den Projektzeitraum zur Verfügung gestellt bekommt, die dann im Projekt verarbeitet werden.

Die Kooperationspartner sind dabei:

- Iunera Verwaltungs-GmbH
- Kahlgrund-Verkehrs-Gesellschaft mbH
- Barnimer Busgesellschaft mbH Eberswalde
- Stadtwerke Oberhausen

Generell zeigt das Projekt, dass auch insbesondere Daten von kleineren ÖPNVs homogenisiert, integriert und für Vorhersagen verwendet werden können. Da man davon ausgehen kann, dass sich die Datenmodelle widerholen, ergeben sich Kostenreduktionseffekte bei der Integration von vielen kleineren ÖPNVs.

Das Projekt zeigt jedoch für iunera auch, dass eine selbsttragende Finanzierung mit drei kleineren ÖPNV-Partnern derzeit nicht mehr initial möglich ist, da neben einmaligen Entwicklungskosten auch die Wartungs- und Betreiberkosten sichergestellt sein müssen um die notwendigen Ressourcen in Personal aufzubauen, da kompetente Fachkräfte im Softwarebereich weltweit stark gefragt sind.

6. Zielgegenüberstellung

Auflagen	**Durchgeführte verwertende Aktivitäten**
Wir entwickeln Methoden in Programmquelltext und wenden offenbarte Verfahren an, um Fahrpläne automatisiert Belegungs-Sensordatenzuzuordnen. Hierzu werden Quelltext Artefakte und Verfahren getestet undangewandt, wie Fahrpläne Sensordaten einander zugeordnet werden können. Hierbei soll vor allem ein universelles intelligentes Verfahren und dessen Einsatz evaluiert werden.	• Verfahren in Quelltext umgesetzt und Quelltext veröffentlicht. • Grundlegendes Verfahren unter anderem vorgestellt auf BAI.CON 2023 - Bavarian Artificial Intelligence Conference Ingolstadt (TH Ingoldstadt Themenbereich ÖPNV)
Entwicklung von Vorhersagen mit statistischer Histogrammbasierter Regression, nach denen zugeordnete Daten transformiert und aufbereitet und danach in einer Time Series Database gespeichert werden, wodurch statistische Abfragen nahe der Echtzeit möglich werden.	• Time Series Database Druid, Konfiguration, Datenmodell und Abfragen mit statistischer Histogrammbasierter Regression als Quelltext veröffentlicht • Vorhersage Microservice als Quelltext veröffentlicht • Grundlegendes Verfahren vorgestellt MLCon München 2023
Entwicklung von Verfahren für Vorhersagen von Belegungsdaten aufgrund adaptiver	• Verfahren in Quelltext umgesetzt und Quelltext veröffentlicht.

Künstlicher Intelligenz und Veröffentlichung der Verfahren.	• Grundlegendes Verfahren vorgestellt MLCon München 2023/2024
Die Ergebnisse des Projektes sind als kostengünstige Software as a Service an ÖPNVs zu vermarkten und zu verkaufen und entsprechend ist eine Gründung einer Servicegesellschaft speziell für den Vertrieb und Betrieb der Software as a Service Plattform anzustreben. Bestrebungen für die Vermarktung sind bereits während der Projektlaufzeit in Kooperation mit Partner zu initiieren (inkl. Erhebung von Bedarfen und Anforderungen).	• Wirtschaftliches potenziell selbsttragendes Angebot ist 2€ pro Bus pro Tag bei Abnahme von für minimum 600 Busse für 5 Jahre. • Zusätzliches Angebot durch die „Open Compensation Token License“ basierte Lizenz mit der jeder die Software offen weiterentwickeln und durch eine faire Anteilsvergütung weiterlizenzieren kann.
Sie verpflichten sich zur Veröffentlichung von Zähldaten, die Linien zugeordnet sind und zur Veröffentlichung über das Verfahren auf dem Firmenblog Ihres Unternehmens. Die Veröffentlichung erfolgt in englischer Sprache, um europaweit lesbar zu sein.	• Software und zugehörige Verfahren ist frei durch „Open Compensation Token License“ nutzbar. Um diese Offenheit möglich zu machen wurde zudem darauf verzichtet den Patentschutz weiter zu verfolgen. • Veredelte Zähldaten aus Testzwecken wurden nicht veröffentlicht, da die Daten und Vorhersagen nach dem KI Act der EU klar den ÖPNVs gehören und wir darüber nicht bestimmen können und uns die Daten rein für das Projekt überlassen wurden.
Sie verpflichten sich zur Veröffentlichung und Diskussion der Projektergebnisse im Rahmen von Konferenzbeiträgen auf mind. 2 verschiedenen Kongressen und Messen. (z.B. Big Data Europe (Vilnius), Mind Mastering Machines (Karlsruhe), Ml Conference (München), TWDI)	Konferenzbeiträge: • BAI.CON 2023 - Bavarian Artificial Intelligence Conference Ingolstadt (TH Ingoldstadt Themenbereich ÖPNV) • Mind Mastering Machines 2023 (KI Konferenz von Heise) • ML Conference 2023 (Machine Learning Conference München) • ML Conference 2024 (Machine Learning Conference München) • Big Data Europe 2023

7. Mittelverwendung und Angemessenheit

Die meisten Kosten entstehen, um Personal zur Forschung und Entwicklung einzusetzen.

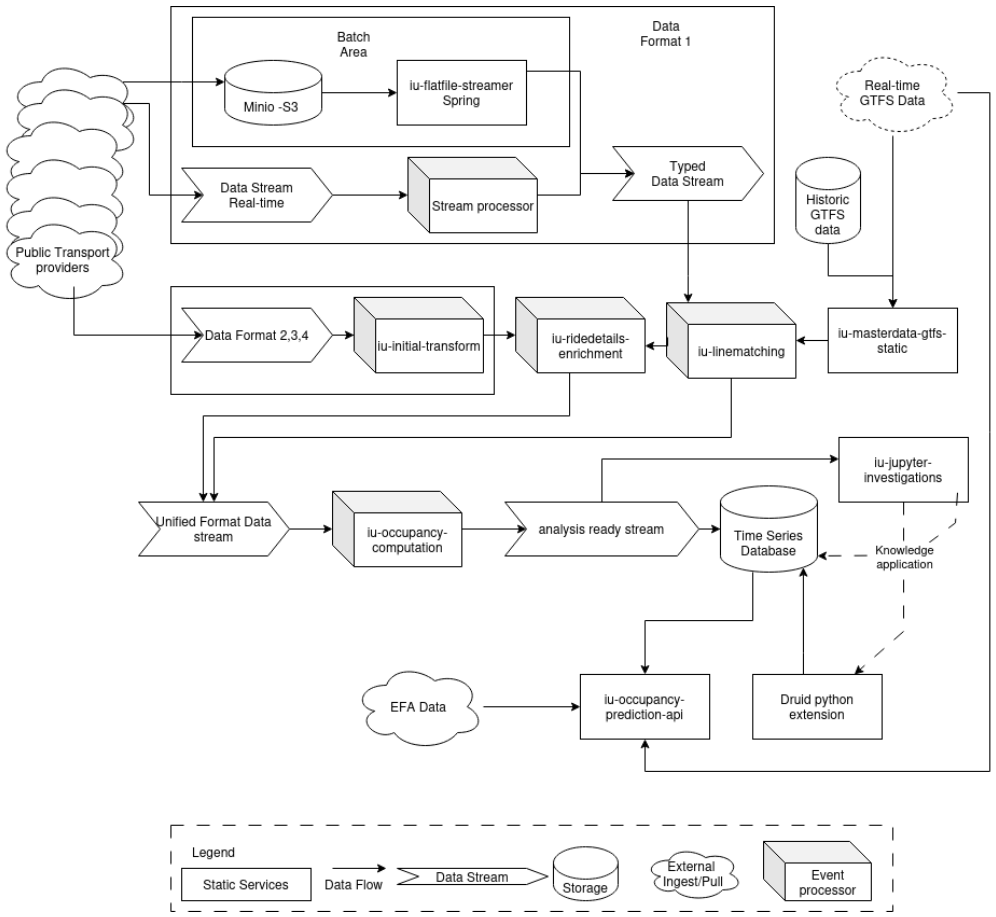
Wie in nahezu allen Künstliche Intelligenz Projekten entsteht der Hauptaufwand bei der Datenverarbeitung, Infrastruktur, Datenvorverarbeitung und bei der Codierung. Dadurch ist das Arbeitspaket, um Daten zu bereinigen geringfügig mehr Aufwand.

Ebenso stiegen unerwartet unsere Aufwände Dokumentationen zu erstellen und Quelltexte in ein veröffentlichbare Form zu bringen. Weiterhin wurde der letztendliche Aufwand für den Endbericht des Projektes im Projekt nicht ausreichend eingeplant. Glücklicherweise konnten diese Mehraufwände großteils durch unsere Schwesterfirma iunera GmbH & Co KG getragen werden.

Weiterhin schaffen wir durch die Reduktion von Reisekosten (getragen durch iunera GmbH & Co KG und remote Teilnahme) und Serverkosten freie Mittel, die somit besser in die Entwicklung einfließen konnten.

Die Reduktion der Serverkosten ist zudem erwähnenswert, da durch den Künstliche Intelligenz Hype oftmals, spezielle hochpreisige GPU-Architekturen verwendet werden (insbesondere Nvidia), die für künstliche Intelligenz vonnöten sind. Durch unseren Ansatz traditionelles Machine Learning zu nutzen konnten wir eine konventionelle Infrastruktur verwenden, diese optimal skalieren und dadurch die Kosten reduzieren. So haben wir, beispielsweise einen eigenen Cloud Cluster aufgesetzt und multiple Services auf einem Knoten zugleich installiert. Durch eine intelligente Partitionierung konnten wir somit horizontal optimal skalieren. Das notwendige Infrastruktursetup hierzu ist zudem als Projektbestandteil in Code veröffentlicht.

Die Gesamtheit und Komplexität des Projektes zeigen sich in der folgenden Übersicht.
Die nachfolgende Übersicht visualisiert die wichtigsten Dienste die konfiguriert, entwickelt und zusammen spielen mussten, um das Projekt zu betreiben. Nahezu alle Dienste und notwendigen Konfigurationen, die in dem schematischen Bild zu sehen sind, wurden letztlich veröffentlicht:



Architektur des Fahrbar20 Projektes. Detaillierte Beschreibungen der wichtigsten Dienste befinden sich in 8. Nutzen und Veröffentlichungen

Aufgrund der großen Anzahl von Konfigurationen und Dienste stellt es unerwartete Aufwände dar, die im Projekt entwickelten Artefakte zu veröffentlichen. Weiterhin ist die Veröffentlichung unter Apache Lizenz bezüglich einer selbsttragenden Verwertung herausfordernd.

Daher stellte sich die zusätzliche und ungeplante Aufgabe parallel zu dem Projekt eine Lizenz zu entwerfen, die diese Rahmenbedingungen beachtet. Die Aufwände einen Lizenzentwurf zu erstellen wurden dabei unentgeltlich dem Projekt zur Verfügung gestellt, da dies nicht in der Projektplanung vorgesehen war.

Diese Zusatzaufwände unterstreichen somit zusätzlich die Angemessenheit der Arbeit im Projekt. Weiterhin zeigt die Vielzahl der veröffentlichen Artefakte (in Code und Konferenzbeiträgen), dass die im Projekt erzeugten Erkenntnisse eine sehr vorteilhafte Quote von Veröffentlichung und Artefakten in Relation zur Förderungshöhe zeigen.

8. Innovative Verwertungsoffensive durch Blockchain basierte Lizenzierung

Unserer Ansicht nach muss die föderalistische Struktur und der Umstand, dass oftmals konkurrierende Eigenentwicklungen in verschiedenen ÖPNVs entwickelt werden in einer Lizenzierung von Software abgebildet werden. Daher wurde auf eigene Kosten, unabhängig vom Projekt, eine Blockchainbasierte Lizenz entwickelt, die die folgenden Probleme adressiert:

- Open Source Software benötigt einen monetären Anreiz zur Weiterentwicklung und zur Instandhaltung. Entwickler, die ein Projekt duplizieren und ggfs. weiter Entwicklern müssen eine Teilhabe gemäß dem Anteil ihrer Weiterentwicklung als Vergütung erhalten können.
- Urheber von Open Source müssen bei der Nutzung ihres Quelltextes fair vergütet werden, wenn Dritte Diesen weiterentwickeln oder Diesen Vertreiben. Dies sichert ein Interesse der Entwickler am Fortbestand von Projekten.
- Die Vergütungsaufteilung muss Blockchainbasiert sein, weil dies der dezentralen Entwicklung entspricht und dies Transparenz und Offenheit schafft.

Die durch das Fahrbar Projekt motivierte „Open Compensation Token License“ ist das Resultat dieser und weiterer Anforderungen:

Ziel der Open Compensation Token License ist es, Entwicklern die Möglichkeit zu geben Open Source zu nutzen und weiterzuentwickeln und für ihre Arbeit eine faire Lizenzgebühr zu verdienen.
https://github.com/open-compensation-token-license/license/blob/main/LICENSE.txt

Im Wesentlichen funktioniert die Lizenz wie folgt:

(1) Entwickler registrieren ihre anteilige Arbeit und deren Abhängigkeiten in der Blockchain durch nicht fungible Token.
(2) Nutzer der Software beschaffen eine Anwendungslizenz für die genutzten Softwareartefakte direkt durch einen Smart Contract von der Blockchain und erhalten hierfür einen nicht fungiblen Token, der die Anwendungslizenz repräsentiert.

(3) Der Umsatz für die Anwendungslizenz wird arbeitsanteilig an die Entwickler verteilt.

Smart Contract basierte Implementierung und Anwendungsdetails sind zu finden unter diesem Hyperlink
https://github.com/open-compensation-token-license/octl

Konkreter Lizenz Mehrwert für das Fahrbar 20 Projekt
Fahrbar und sämtliche zugehörige Artefakte sind unter der „Open Compensation Token License“ verfügbar.

Dies bedeutet, dass hiermit ein virtuelles Projektkonsortium zur Weiterführung geschaffen wird, indem jeder Entwickler und ÖPNV die Software selbst weiterentwickeln oder Anwendungslizenzen erwerben kann.

Somit ist eine Verwertung und offene Weiterentwicklung von Fahrbar 20 lizenzbasiert möglich.

8. Nutzen und Veröffentlichungen

Der Nutzen der Projektergebnisse wird im Folgenden durch die Beiträge der einzelnen veröffentlichen Artefakte dargelegt:

8.1 Linienzuordnung zu GPS Daten

Die Umsetzung des Verfahrens zur Linienzuordnung wurde mittels einen Datenstromprozessors (Stream-Prozessors) auf Grundlage von Apache Flink durchgeführt. Konkret nutzt der Stream-Prozessor GTFS-Daten aus einem GTFS-Microservice für Haltestellen und Linien und verarbeitet dann einen Datenstrom, um die Datenpunkte für Linie und Haltestelle zu kennzeichnen. Grundsätzlich basiert der Abgleich auf der Nähe von geografischem Standort und Zeit für mehrere Haltestellen und deren Zeiten, um die wahrscheinlichste Linie/Route und Haltestelle für jeden Datenpunkt abzuschätzen.

Weitere beschreibende Details und Quelltext: https://github.com/iunera/iu-linematching

Die entwickelten Methoden funktionieren in den meisten Fällen korrekt und ermöglichen Tradeoffs zwischen Erkennungsgeschwindigkeit und Genauigkeit. Das heißt, dass je schneller erkannt wird desto eher die Wahrscheinlichkeit, dass die Zuordnung falsch ist und umgekehrt. Dies ist in den meisten Fällen jedoch unerheblich, da der Takt des ÖPNVs auf dem Land nicht sehr hoch ist und die Daten oftmals nicht in Echtzeit vorliegen.

Interessanterweise haben auch unsere Tests gezeigt, dass Fahrwege (d.h. welche Straßen ein Bus wann entlangfährt) zu sehr schnellen und guten Erkenntnissen führen allerdings diese Fahrwege von den meisten ÖPNVs nicht veröffentlicht werden. Die Rekonstruktion von Fahrwegen durch eine Vielzahl von GPS-Daten ist jedoch durch Big Data Technologie trivial möglich.

Die bedeutet somit:

- Es ist empfehlenswert, dass politische Rahmenbedingungen geschaffen werden damit konkrete Fahrwegen neben Liniendaten veröffentlicht werden.
- ÖPNV-Fahrwege können recht problemlos durch Institutionen, die Zugriff auf eine Vielzahl von Smartphone GPS Daten haben (wie z.B. Technologie Konzerne oder Smartphone Hersteller) rekonstruiert werden. Dies könnte beispielsweise dann dazu genutzt werden das solche Konzerne aufdecken welche Busse wann besonders voll sind. Für kritische Infrastruktur ist dies eine potenzielle Sicherheitsproblematik.

8.2 Weiterentwicklungen zu Apache Flink

Da unser Szenario nicht komplett durch Apache Flink Bordmittel abbildbar ist, entwickeln und veröffentlichen wir verschiedene Erweiterungen für das Framework. Die wichtigsten hiervon und deren Mehrwerte werden nachfolgend beschrieben.

Erweiterungen von Verarbeitungsfenstern
Flink unterstützt Ereigniszeit-oder Echtzeit-Verarbeitungszeitfenster. Deren Nutzung beruht auf einem logischen ODER, was bedeutet, dass man sich auf die Ereigniszeit oder die Bearbeitungszeit einstellen kann. Ereigniszeit bedeutet, dass es keine Rolle spielt, wie viel Echtzeit zwischen den Daten verstrichen ist, solange kein Element eintrifft. Echtzeit bedeutet, dass rein die verstrichene Echtzeit zur Verarbeitung ausschlaggebend ist. Dies kann problematisch sein, wenn man die Ereignisdatenpakete aggregieren/gruppieren möchte und die Aggregation ausgelöst werden soll, wenn eine Minute lang kein Datenpaket eintrifft, da somit gewartet wird, bis das nächste Datenpaket eintrifft – auch wenn dies erst am nächsten Tag der Fall sein sollte. Dies ist gerade bei ÖPNV-Daten relevant, weil die Daten oftmals irregulär übermittelt werden.
Wir bieten daher eine spezielle Art einer Windowing Funktion an, die es ermöglicht Echtzeit und Ereigniszeit zugleich zur Aggregationsauslösung zu nutzen.

Geotrigger für Verarbeitungsfenster
Um schnell massive Datenreduktionen durchzuführen ist es auch vonnöten Daten auf Geopositionen aggregieren zu können. Daher haben wir einen Geotrigger entwickelt mit dem man performant bei einer sich verändernden Geoposition, in gewünschter Granularität, Aggregationen auslösen kann.

Weiterhin befinden sich noch eine Helferfunktion und deren genaue Beschreibung mitsamt Quelltext unter: https://github.com/iunera/iu-flink-extensions

8.3 GTFS Service

Um GTFS-Daten flexibel importieren und nutzen zu können wird eine GTFS Kompetente benötigt.

Hierfür wird ein Microservice zum Verarbeiten statischer GTFS-Daten in einer PostgreSQL-Datenbank bereitgestellt. Das Datenmodell verfügt über die Erweiterung, dass eine GTFS-Datei in verschiedenen Revisionen gespeichert werden kann, was die Extraktion historischer Zeitpläne ermöglicht.

Der Hauptgrund für die Entwicklung dieses Dienstes war, dass wir keine gebrauchsfertige GTFS-Software finden konnten, um GTFS-Dateien einfach aus Ordnern oder URLs zu

importieren und Abfragen mit geobasierten und/oder anbieterbasierten Slicing durchzuführen. Darüber hinaus benötigten wir historische GTFS-Daten, in denen verschiedene Anbieter und verschiedene Revisionen derselben Daten gleichzeitig gespeichert werden können.

Durch die Speicherung verschiedener Metadatenschlüssel im GTFS-Modell kann man mit diesem Dienst nun Folgendes durchführen:

- Dynamisches Erzeugen von einzigartigen Zeit-Geoschlüsseln, die einen Halt einzigartig kennzeichnen
- Abfrage historischer Daten
- Vergleichen von GTFS-Datenänderungen
- Zusammenführung von Daten verschiedener Anbieter und Neuaufteilung
- Partitionierung aufgrund von Geokoordinaten, Historie, Importvorgang oder Anbieter

Weiterhin befinden sich noch Helferfunktionen und deren genaue Beschreibung mitsamt Quelltext unter: https://github.com/iunera/iu-masterdata-gtfs-static

8.4 Datentypen und Schlüsselgenerierung

Um Daten aus unterschiedlichen Datenquellen abzugleichen benötigt es entsprechende Datenmodelle und Zuordnungsmechanismen.

Unsere Lösungen für diese Herausforderungen sind im Wesentlichen:

- Domänenobjekte für den öffentlichen Verkehr
- Ein universeller Schlüsselgenerator (Fuzzy-Hashing-Technik) zum Abgleichen von Fahrplänen sowie Fahrten und Haltestellen aus verschiedenen Datenquellen

Schlüsselgenerator
Der Schlüsselgenerator basiert auf lokalem sensiblem Hashing und verschiedenen Datenclustering-Techniken. Die Idee besteht darin, Kollisionen innerhalb unterschiedlicher Zeit- und Geogranularität zu maximieren, um sie an verschiedene Datenquellen anzupassen.

Domänenobjekte
Die Domänenobjekte orientieren sich an den Datentypen https://schema.org, spiegeln die Internationalisierung wider und modellieren Datentypen im ÖPNV.

Die genaue Beschreibung mitsamt Quelltext ist zu finden unter: https://github.com/iunera/iu-transport-datatypes

8.5 Apache Druid Erweiterungen

Um Schnelltests mit Apache Druid durchzuführen, entwickeln wir Druid Erweiterungen.

Besonders erwähnenswert ist der Druidimporter, welcher es ermöglicht Python Quelltext, der in der Data Science Community üblicherweise vorherrschend ist, in Apache Druid Ingestions einzusetzen.

Die genaue Beschreibung mitsamt Quelltext ist zu finden unter:
https://github.com/iunera/iu-code-ingestion-druid-extension

https://github.com/iunera/druid-exporter

8.6. Daten Anreicherung

Wir haben in unseren Datenanalysen gesehen, dass sich Datenanreicherungen anbieten um Vorhersagen zuverlässiger zu gestalten. Daher entwickeln wir Prozessoren um Daten dynamisch und schnell anzureichern.

Diese Anreicherungswerkzeuge berechnen im Wesentlichen:

- Die absolute Veränderung der Passagiere
- Die absolute Höhe des Passagierwachstums
- Die Gesamtstoppdauer
- Die Abfahrtsverzögerung
- Die Ankunftsverzögerung
- Eintrittsgeschwindigkeit der Passagiere
- Passagierausstiegsgeschwindigkeit
- Wie lange waren wie viele Passagiere zusammen in einer Fahrt
- Natürlicher Schlüssel der Haltestelle
- Schulferien
- Stopfunktion (z.B. ist ein Bahnhof oder eine Schule in der Nähe)

Die genaue Beschreibung mitsamt Quelltext ist zu finden unter:
https://github.com/iunera/iu-ridedetails-enrichment

8.7 Daten Streamifizierer

Damit Daten homogen verarbeitet werden können, werden Zähldaten, die in statischen Dateien geliefert werden „Streamifiziert“. Somit können alle Daten zukunftssicher, homogen als Datenströme verarbeitet werden und Aufwände durch eine Unterscheidung zwischen verschiedenen Arten der Datenlieferung werden minimiert. Hierfür bieten wir einen Microservice an. Dieser liest Personenzählerdateien mit unterschiedlichen Datenformaten ein und wandelt sie in einen Datenstrom um.

Die genaue Beschreibung mitsamt Quelltext ist zu finden unter:
https://github.com/iunera/iu-flatfile-streamer

8.8 Infrastruktur als Code

Ein wesentlicher Teil unserer Arbeit besteht auch darin eine cloudagnostische Infrastruktur zu betreiben auf welcher sämtliche Komponenten lauffähig sind.

Daher werden die Resultate wie man eine solche Infrastruktur erzeugt in verschiedenen Projekten veröffentlicht:

https://github.com/iunera/druid-cluster-config
https://github.com/iunera/jupyterhub-cluster-config
https://github.com/iunera/kafka-cluster-config

https://github.com/iunera/mongodb-cluster-config
https://github.com/iunera/helm-charts
https://github.com/iunera/azure-devops-templates

Genauere technische Beschreibungen und der Zweck der einzelnen Komponenten befinden sich dabei in den verschiedenen Quellcodeprojekten unter den genannten Links.

8.9 Konferenzbeiträge

- **BAI.CON 2023** - Bavarian Artificial Intelligence Conference Ingolstadt (TH Ingoldstadt Themenbereich ÖPNV)

  Die AIMotion ist eine Reihe von Fachkonferenzen zu Künstlicher Intelligenz in Bayern. Die BAI.CON zieht insbesondere Forscher an.

- **Mind Mastering Machines 2023** (KI Konferenz von Heise)

  Die Minds Mastering Machines ist der Treffpunkt für Fachleute, die Machine Learning und KI in die technische Realität umsetzen.

- **ML Conference 2023** (Machine Learning Conference München)
- **ML Conference 2024** (Machine Learning Conference München)

  MLCon bringt Experten und Enthusiasten für einen viertägigen intensiven Einblick in die neuesten Trends, Algorithmen und Anwendungen im Bereich ML und KI zusammen.

- **Big Data Europe 2023**

  Die Konferenz umfasst eine Reihe von Workshops und Sitzungen, die darauf abzielen, Entwickler, IT-Experten und Benutzer zusammenzubringen, um ihre Erfahrungen auszutauschen, Best Practices zu diskutieren und Anwendungsfälle und Geschäftsanwendungen im Zusammenhang mit ihren Erfolgen zu beschreiben. Die Veranstaltung soll aufklären, informieren und inspirieren.

9. Entwicklungen während der Durchführung des Vorhabens

Generell ist die Entwicklung, die unserer Meinung nach, das Projekt und den ÖPNV am meisten betrifft die Zunahme der geopolitischen Instabilität. Der ÖPNV ist Teil der kritischen Infrastruktur und es ist die Frage zu stellen, in wie fern Open Source Software und offene Daten bei kritischer Infrastruktur generell eine Zukunft haben.

Dieser Umstand sollte unseres Erachtens durch sicherheitskritische Betrachtungsweisen speziell im ÖPNV weiter untersucht werden, da dies starke Auswirkungen auf die

Entwicklung und Kommerzialisierung von Entwicklungen und zugehörige Geschäftsmodelle hat.

10. Fazit

Ein Hauptproblem in der Verkehrswende ist die Verfügbarkeit von Daten und die Fragmentierung des ÖPNV. Unser Projekt zeigt, dass es möglich ist, quellcodeoffen ÖPNV übergreifende Software zu entwickeln, mit welcher Belegungsanalysen und Belegungsprognosen möglich sind.

Mehr denn je glauben wir nach diesem Projekt daran, dass die Politik gefragt ist im ÖPNV gesetzliche Rahmenbedingungen durch Datengrundlagen zu schaffen. Solche Rahmenbedingungen bedeuten in der Praxis konkret, dass weitaus mehr Zähldaten als heute, Verwendungszweck ungebunden und verfügbar sein müssen. Dies würde es ermöglichen neue Geschäftsmodelle durch die Daten zu schaffen und dogmafreie Analysen und Bestandsaufnahmen durchzuführen.

Aus technischer und Projektsicht sind wir der Überzeugung, dass unsere Entwicklungen und insbesondere die innovative Blockchainbasierte Lizenz ein Zukunftsweg sein kann, wie vor allem im ÖPNV und Open Source Bereich Innovationen entstehen können. Denn dies ermöglicht besonders fragmentierten ÖPNV Bereich Co-Innovationen und Entwicklungen.

Unsere nächsten Schritte sehen wir darin für unsere Entwicklungen und für das Blockchainbasierte Lizenzmodel Feedback zu erhalten. Sobald sich hierfür klare finanzielle Zusagen im ÖPNV-Umfeld abzeichnen werden wir das Projekt in die nächste Phase führen.