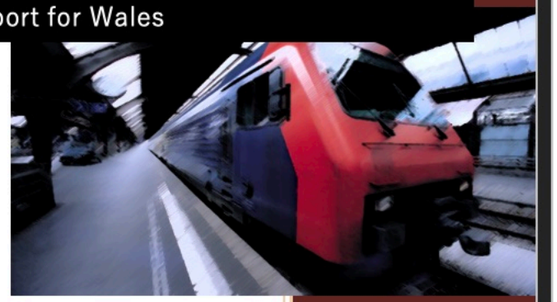
Proposed Database Design for Transport for Wales



KTK Consultants

10/31/2022

Table of Contents

Proposed Database Design for Transport for Wales

By KTK Consultants

Introduction

In 2021, the Ministry of Transport and Economy for Wales launched the Llwybr Newydd – the Wales Transport Strategy 2021. This 20-year strategy aims to reshape the transport system and focuses on people and climate.

KTK Consultants (KTK) have been commissioned to design a logical database that will be the backbone of the integrated ticketing system. This system will allow passengers to purchase tickets and use them across all modes of transport provided by Transport for Wales (TFW) (one card, all services). KTK is pleased to present a proposal design report for the Transport for Wales (TFW) data management system.

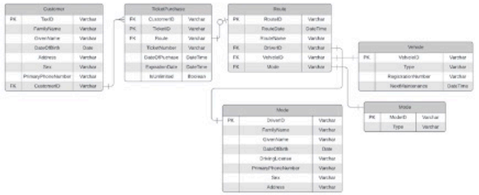
Logical Design

The logical design showcasing the data points, entities, attributes and relationships is presented below. It must be noted that the logical design presented in this section follows the database schema, but it serves only as a high-level presentation of the data overview and is meant to be human-friendly. See the section "Database Design" for the technical details.

Purpose Specification

The features of the developed system are explained below:
-***Customer management***: This feature collects and holds personal information about the customers, aiming to personalized ads, better customer relationship management, transaction tracking and fraud prevention;
-***Reservation management:*** Reservation management allows the handling of specific routes, mainly for fixed track means of transport with numbered seats. Indicatively the feature allows the management of date, time, point of departure, point of arrival and the mode of transport;
- ***Availability management:*** Tracking the availability of specific routes for specific times and dates, allowing the reservation of available seats;
-***Ticket management:*** It involves identification information for the tickets purchased by the customers, such as the ticket number and the ticket ID;
-***Purchase channel:*** This feature specifies the method of booking the ticket, being online, through the app or the web, or through one of the certified agents.

Entity Relationship Definition



The data will be pushed in a XML format during the transactions between the systems (embedded sensors, front-end applications) to the database. The format has been selected due to its reusability, extensibility and ease of parsing (IBM, 2021). Standard and simple data types have been used with varchar between variable length strings (texts) and Date and Date Times being ISO formatted date or date time values in UTC (e.g. 2022-10-31T17:44:05+00:00) (IONOS, 2020).

Data Management Pipeline

The integrated ticketing system will allow passengers to purchase tickets or load their travel cards; these tickets and cards can be used to access any mode of transport service provided by Transport for Wales. Customers should be able to purchase tickets across the various sales channels offered by Transport for Wales, mainly through the mobile App, on the official website, at the various transport stations, and at selected retail outlets.

In addition, TFW would want to analyze information related to popular service routes, sales by mode of transport, and busiest routes by the hour, week, or month. For this purpose, we propose a data management pipeline that collects data from multiple sources, transforms it and consolidates it into one data storage hub. KTK proposes a combination of batch processing and streaming processing pipelines to take into account the time-sensitive nature of the data being collected. The streaming processing pipeline allows TFW to receive up-to-date information on ticket sales and fleet movements and forecast the demand for transport services along routes, while batch processing will be essential for analyzing historical data to understand medium to long-term trends (Alexeyl, 2022).

In order to achieve this, three crucial steps must be considered and undertaken. These are identifying the source of the data, the processing steps, and the destination of the processed data.

Data Source

For the integrated ticketing system, data will be sourced from IoT device sensors located at train stations, in buses, and at bicycle sharing stations, the Customer Relationship Management (CRM) system that processes ticket purchases, and the Enterprise Resource Management system which manages the fleet, accounting, and supply chain for Transport for Wales.

Processing

The processing stage is where data cleansing will occur. This involves error detection and repairing of detected errors. During this stage, we propose the use of data-cleaning pipelines, which use automated data-cleaning libraries (Krishan & Wu, 2019). Another alternative method would be to generate small random samples of data and explore the statistical properties, and set cleaning rules. Saluyan (2019) describes Random Sample Partitions (RSP) which is a distributed data model used to represent a large dataset through ready-to-use sample data blocks. Using this method, we can identify potential types of value errors, understand the data and retrieve samples of clean data.

Destination

For this logical database design KTK proposes using both a data warehouse and a data lake to store data. The data warehouse will house historic data that are summarized. The data lake will store all the unstructured data such as data from the IoT sensors, files or even video footage captured at bus and train stations for example. A data lake will allow TFW to store large amounts of data which can be used in machine learning and artificial intelligence at a later stage.

Database Design

The Entity Relationship Diagram (ERD) following below is already adapted to conform with the database design normal forms, up to BCNF (Date, 2019).

1NF

For a table to be in the 1st normal form, every value should be atomic, each record needs to be unique, values should be of the same domain and all the columns per each table should have unique names.

2NF

For a table to be in the 2nd normal form it should be already in 1st normal form and also have no partial dependency.

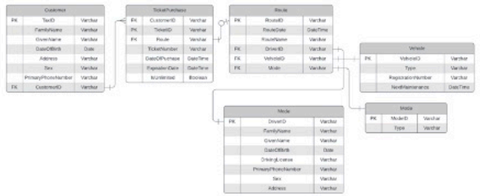
3NF

In order for a database to be in the 3rd normal form it should be in the 2nd normal form and also have no transitive dependencies. Transitive dependencies are found when a 'non-prime' attribute depends on other non-prime attributes rather than the prime attributes or prime key' (StudyTonight, no date) - if the prime key is composite. All the tables therefore are in the 3rd normal form.

BCNF

In order for a database to be in Boyce-Codd Normal Form (BCNF) it needs to be in third normal form and also for any dependency x to y, the x cannot be a non-prime attribute if y is a prime attribute (StudyTonight, no date).

ERD



Summary

The needs of Transport for Wales are dynamic and as such the single logical database design aims to incorporate this in the proposed design.

References

1. Alexsoft. (n.d.). *What is Data Pipeline: Components, Types, and Use Cases*. [online] Available at: https://www.altexsoft.com/blog/data-pipeline-components-and-types/ [Accessed 29 October 2022].
2. Krishnan, S. and Wu, E., 2019. Alpaclean: Automatic generation of data cleaning pipelines. arXiv preprint arXiv:1904.11827.
3. Llwybr Newydd (2021) The Wales Transport Strategy 2021. Available from: https://gov.wales/llwybr-newydd-wales-transport-strategy-2021 [Accessed 27 October 2022].
4. Neutatz, F., Chen, B., Abedjan, Z. and Wu, E., 2021. From Cleaning before ML to Cleaning for ML. *IEEE Data Eng. Bull.*, *44*(1), pp.24-41.
5. Saluyan, S., Huang, J.Z. & He, Y., (2019). Exploring and cleaning big data with random sample data blocks. *Journal of Big Data* 6(45). Available at: https://doi.org/10.1186/s40537-019-0205-4. Accessed: [30 October 2022].
6. IBM. (2021) Uses of XML. Available from: https://www.ibm.com/docs/en/i/7.2?topic=introduction-uses-xml [Accessed 31 October 2022].
7. IONOS. (2020) ISO 8601 – Effectively Communicate Dates and Times Internationally. Available from: https://www.ionos.com/digitalguide/websites/web-development/iso-8601/ [Accessed 31 October 2022].
8. Date, C. J. (2019) *Database Design and Relational Theory: Normal Forms and All that Jazz*. California: apress. Available from: https://tinyurl.com/25epdfk7 [Accessed 31 October 2022].
9. StudyTonight. (no date) Third Normal Form (3NF). Available from: https://www.studytonight.com/dbms/third-normal-form.php [Accessed 31 October 2022]

- Knowledge and under-standing of the topic/ issues under consideration The report offers a good demonstration of knowledge and understanding of relevant topics. The sections further include very good understanding in the relevant key areas of knowledge. The logical database design includes purpose specifications and entity relationship definitions. The data managemnet pipeline discusses data sources, processing, and destination. The database design normal forms are briefly indicated. The reference section includes nine references which are cited to support arguments in various sections. Application of knowledge and understanding The project redresses a practical case of designing a database for an integrated ticketing system commissioned by a transport authority. Theoretical descriptions are linked to the practical requirements as communicated by an authority's strategy document. The database design section could have been more elaborate and meaningful. Criticality There was a satisfactory demonstration of critical analysis with regards to linking theory and practice. This is particularly apparent in the section on data management pipeline. The discussion could have been more focused, for example the data sources being more specific on information systems access. Structure and Presentation (as detailed in the assessment guidance) The word count is around 1000 +10% (excluding TOC and references), and hence within the limit. The report structure follows the requirements as it includes three sections on logical design, critical evaluation of the data pipeline, and database proposal. Although not affecting the word count, the ERD diagram is included twice with the latter (in the data design section) being redundant and could have been replaced with more elaborate explanations. The conclusions section is very short and seems to have been included just for the sake of having a conclusion section.