Data Storage and GDPR Compliance Strategy: Practical Guide

Individual Project: FestivalConnect

Name: Lucas Jacobs
Class: S-A-RB06
PCN: 490692
Student number: 4607368
Technical teachers: Felipe Ebert, Bartosz Paszkowski
Semester coach: Gerard Elbers

Table of Contents

Glossary

Term	Definition
Persistent Data	Persistent data in the field of data processing denotes information that is infrequently accessed and not likely to be modified.
Data replication	The process of copying data from one location to another.
Distributed system	A collection of computer programs that utilize computational resources across multiple, separate computation nodes to achieve a common, shared goal.
Nodes	Individual Computers or servers within a distributed system. Node can store and process data independently.

Introduction

This document will cover the importance of how data is being stored in FestivalConnect, and how this is securely done, with the consideration of GDPR. With the General Data Protection Regulation (GDPR), which his implemented by the European Union in May 2018, protecting human privacy rights is becoming more important. With GDPR it tries to give individuals more and greater control over their data, and with these regulations, FestivalConnect needs to shape their way to handle and process data legally. (General Data Protection Regulation, 2024)
This Report will cover the requirements and examine the key principle of GDPR. We will identify how FestivalConnect is collecting, processing, and managing data and we can ensure that this will comply with GDPR. Furthermore, it will go over how the implementation of technical security measures, data storage, and maintaining data privacy and security standards.

Understanding Data Requirements

In FestivalConnect, there needs to be a proper connection between festival-goers and organizers, which requires proper management of persistent data to make sure that the user is satisfied. Therefore we need to look at the following properties and requirements on persistent data which is important for FestivalConnect.

- **User Profiles**: The essential information of personal details, festival preferences, and community associations are important for personalized experiences and targeted communications.
- **Event Data**: Information regarding the festivals, including data, locations, genres, etc. This will trigger users to discover and engage with events that are upcoming and within their interests.
- **Community management**: Data to festival communities, such as member lists and post feeds, can improve community engagement among users.
- **Posts**: Inside of communities, keeping track of posts and reactions, to maintain a positive user experience.
- **Notification Records**: Keeping track of notifications that are sent to users regarding certain updates or new events, to keep communication and engagement clear.
- **Authentication and Authorization**: Securely storing user's authentication credentials, and proper management of permissions based on roles and privileges for certain keeping accounts and sensitive data secure.
- **Data Privacy**: Following the privacy regulations (GDPR) to get user consent, managing data access and usage, and data subject rights such as gaining access, rectification, and deletion.
- **Scalability and Performance**: Data structures and storage systems that can be easily scaled, to handle peak usage periods in the festival season.
- **Data Integrity and Security**: Design FestivalConnect to take measures to maintain data integrity, prevent unauthorized access, and protect against cyber threats, to be able to maintain user trust.

All these cover important parts of managing persistent data and need to be in line with GDPR so that FestivalConnect can effectively manage persistent data while assuring the privacy and rights of the user. Moreover, important requirements around FestivalConnect can be found in (Jacobs, 2024, Software Requirements Specification).

Privacy-Sensitive Data and GDPR Compliance

FestivalConnect needs to make sure to safeguard and identify privacy-sensitive data, to ensure that it complies with GDPR. When this does not happen, according to article 83, FestivalConnect can receive a fine of up to 20 million or four percent of the yearly turnover. (Art. 83 GDPRGeneral conditions for imposing administrative fines, n.d.)

An important outline of GDPR

the following is important to know regarding the GDPR and what FestivalConnect needs to consider on how to comply with it.

- The user needs to give consent in order to manage and use their data for one or more purposes, also needs to be freely given and revocable. This processing of data needs to be limited to articles 5, 6, and 7.
- Rights: Humans have rights, so they need to know have rights upon the following (Covering 12-22):
    - They need to know that they collected your data
    - What is collected?
    - The right to correct data.
    - Ask to delete the data, or the option to delete certain data.
    - Right to get the data in a structured commonly used and machine-readable format such as JSON, XML, and CSV.
        - Data subjects can request an export in such format from you of their data, this can also be done to ask from your competitors.
    - Rectify your data.
- Security measures to protect data, minimize the processing of only what is needed to be processed of data, keep a record of processing activities and you need to notify authorities within 72 hours in case of a breach. (Art 25, 30, 32, 33)

(Brodwall, 2019)

Sensitive Data and GDPR

When considering privacy-sensitive data, the following is taken into account.

- Health-related data
- Trade-union membership
- Personal information regarding racial or ethnic origin, political opinions, religious or philosophical beliefs.
- Genetic data, biometric data that is processed to uniquely identify a person.
- Data relating to individuals' sex life or sexual orientation.

(What personal data is considered sensitive?, n.d.)

For FestivalConnect, the only privacy-sensitive data that they need to acquire and properly manage is related to personal information.

Personal Information In FestivalConnect

The following data is being stored by users, categorized into which sensitive information this can relate to.

- **User Profiles**: names, and email addresses, but also preferences to genres can correspond to personal information regarding racial or ethnic origin.
- **Community Registration (Event):** Registration on festivals and expressing your opinion in posts can indicate festivals users are interested in attending, and may indirectly reveal religious or philosophical beliefs or even political opinions.
- **Notification Records**: This can be detailed communication sent to users, and may contain personal information. Which can be explicit to health-related data or data concerning an individual sex life or sexual orientation.
    - Example: Health and safety measures at festivals, such as staying hydrated or information about medical facilities onsite, can reveal indirect health-related data about a user.

    - Another thing is an LGBTQ+ Community event. FesivalConnect can also facilitate communities towards LGBTQ+ events, notifications about such events could indirectly reveal information about users' sexual orientation.
- Authentication Credentials: User login information, email, and passwords. Crucial to safeguard this to comply with GDPR.

Functionality Plans to Comly with GDPR

To safely and properly manage this private information, the following requirements are taken into account.

- Data Minimization: Only collect and process the needed data required for specific purposes, to make sure that data is not retained. So FestivalConnect will not hold more personal data than necessary.
- Purpose Limitation: Have a clear purpose for what is being collected and stored, to ensure that data is not used for unrelated activities.
- Consent Management: By giving the user the option to freely and explicitly give consent before processing data, with the option to always revoke consent when needed. The regulation needs to be set to default uncheck.
- There will not be a privacy policy, since this project is a prototype for a real-world application.
- There is no Data Protection Officer, this is an individual project.
- Data Subject Rights: the user can easily access personal data, and update or delete their data.
- Security Measures: Proper security to ensure that data can only be used by the appropriate authorized user, for example, users can only update their data.
- This is a school project with a prototype for the real world, and will not include a breach notification system when there is a data breach, it will notify within 72 hours.

(Bhatia, n.d.)

With these requirements, FestivalConnect makes sure that privacy-sensitive data is protected and complies with GDPR.

Evaluating Candidate Data Stores Using CAP Theorem

CAP theorem can be described as three important requirements that constantly compete against each other in a distributed system with replication. With this tool, FestivalConnect can be aware of the trade-offs in designing the system. The CAP stands for consistency, availability, and partition tolerance. The theorem states that it is not possible to guarantee all three properties at the same time in the distributed system with data replication, instead it is only possible to support two of the three properties.

With consistency, we mean that in a system all the nodes have the same up-to-date data at all times, providing a uniform view of information across transactions. Moreover, availability will guarantee that every request that either reads or writes data is successful or gives a clear view of failure, having each node reposing within a reasonable period. Partition tolerance will allow the system to work, even when the parts are failing, which will make sure that there is consistency despite the network errors. (The CAP Theorem in DBMS, 2023)

To give an overview of which databases consist of what properties, here is a figure to further illustrate.

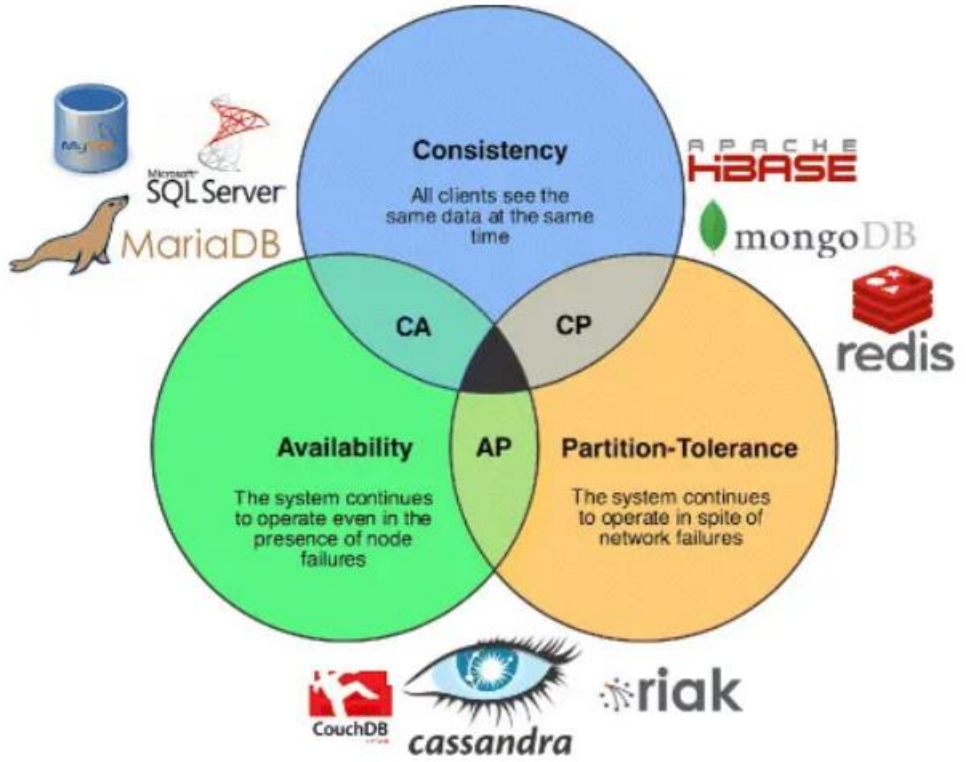

Figure 1: CAP Theorem With Database Examples (Choudhary, 2023)

So for the data requirements plans of FestivalConnect, how can these fit compared to the CAP theorem?

- **Consistency**: FestivalConnect needs to make sure that there is consistency in handling privacy-sensitive data, with the concern of GDPR. This means that we need to maintain up-to-date records of user consent, data usage, and processing activities.
- **Availability**: FestivalConnect guarantees the availability of privacy-sensitive data while complying with the GDPR. This ensures that users can access their data when needed and the system is protected from breaches. Think of implementing a correct authentication mechanism to prevent unauthorized access to personal information.
- **Partition Tolerance**: To ensure data integrity and security that are measures taken to prevent cyber attacks, this can also be further found inside of the requirements (Jacobs, 2024, Security Design). Also for example the database of the posts is hosted on Azure, which will also ensure a single point of failure, which can be horizontally scalable to host the other database on this as well.

To cover CAP, FestivalConnect uses different data storages such as MySQL, Redis, and RabbitMQ, which is essential to cover all the properties: consistency, availability, and partition tolerance. This can also be seen in Figure 1.

- **MySQL (Relation Database):**
  - Consistency: MySQL, offers consistency to ensure that data is always accurate and up to date which is important for handling privacy-sensitive data and GDPR compliance.
  - Availability: MySQL has high availability with features such as replication and failure mechanisms, to ensure to access the database even in node failure.
- **Redis (In-Memory Data Store):**
  - Consistency: The main focus is on performance and low latency (caching), Redis can provide eventual consistency with replication mechanisms.
  - Availability and Partition Tolerance: Redis ensures that the system remains accessible and operational even when network partitions or node failures. Also, Redis supports clustering to improve availability and throughput.
  - Redis is more towards the AP aspect than a CP in the CAP theorem. (Redis , n.d.)
- **RabbitMQ (Message Queue):**
  - Availability: RabbitMQ will allow FestivalConnect to give a queue and deliver messages even when a node fails.
  - Partition Tolerance: RabbitMQ makes sure that it is designed to be partition tolerant, which will benefit FestivalConnect in message processing even if certain parts of the system become not reachable due to network partitions. (Romero, 2018)

All and all, since 2000, many things have been said about the CAP theorem and been criticized for oversimplifying important concepts and ignoring factors such as time delays, which are deeply connected to partitioning in distributed systems. In addition, the binary classifications of CP or AP overlook the real-life scenarios and the trade-offs involved in balancing consistency and availability. When FestivalConnect there distributed system evolved and became more complex, there was a need to look into more factors beyond the CAP theorem to effectively design the system. (Özkan, 2021)

Designing Architecture with Distributed Data Stores

To design an architecture with distributed data stores, several patterns can be illustrated to ensure scalability without costing any performance or reliability, here are the following things that are applied in FestivalConnect.

- Multitier Architecture: FestivalConnect uses microservices, with each service having separate layers such as the controller, logic, and data storage. It is using .net where you can easily separate each layer using libraries. This separation ensures scalability and flexibility in managing different components.
- Data Store Selection: Each service uses its database for distributed data. For databases, MySQL is used and for in-memory data, Redis is used for frequently accessed data.
- Event-Driven Architecture (message-broker): For sending notifications for certain events, RabbitMQ is used, making sure of scalability and fault tolerance.
- Replication and Redundancy: The Master-Slave strategy is being used for the database services. Replication is where the changes that are made on the master database are properly propagated to the slave database to make sure that they stay synchronized. With redundancy, we mean the duplication of data to ensure that operation can continue in case of fault tolerance when something fails. In this case, when the master fails, the slave will serve as the master, keeping the application running and preventing a single point of failure. To give more clarity:

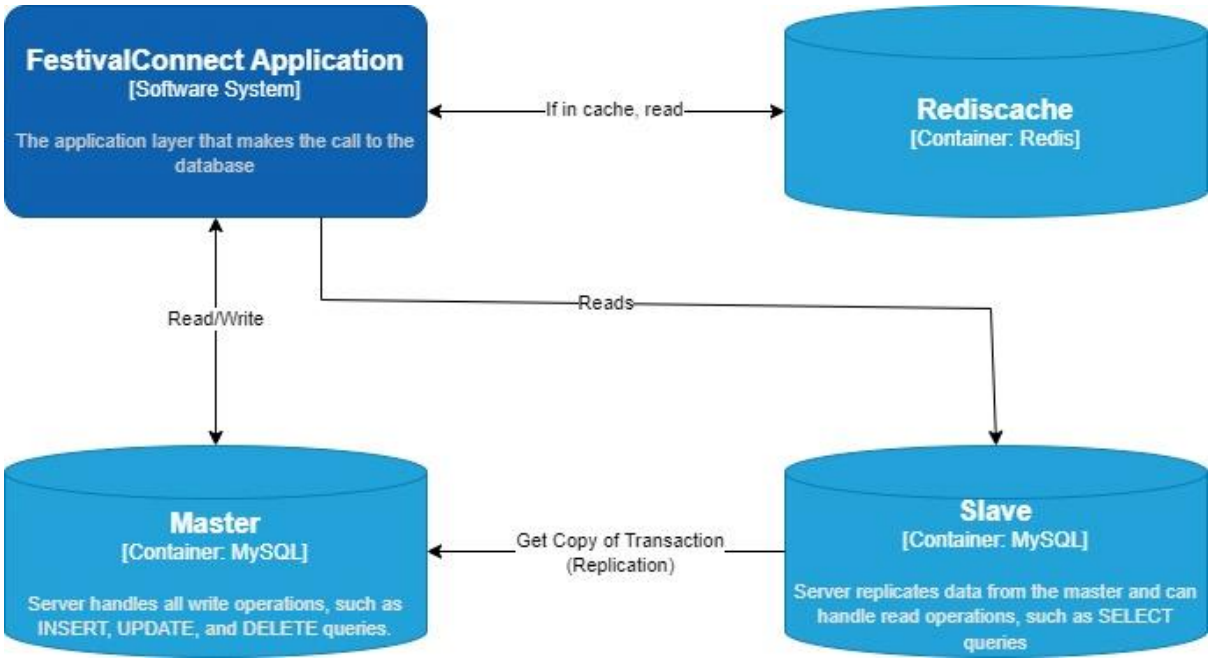

Figure 2: Master-Slave Database Diagram

This can then be horizontally scalable with having multiple slaves which users can read from.

Conclusion

To conclude, FestivalConnect analyzed data storage and GDPR compliance for the application and the outcomes will successfully be incorporated, by implementing several data storages and other approaches to manage the persistent data. The architectural choices with the specific data requirements that FestivalConnect has set, ensure scalability, performance, and data integrity and it is compliant with GDPR. Furthermore, with storages of distributed data using MySQL, Redis, and RabbitMQ, FestivalConnect covers the key aspects of CAP theorem and with the architectural decisions of multitier and event-driven, will ensure scalability and fault tolerance. This design will FestivalConnect have an application that has data management and privacy protection that aligns with expected industry standards.

References

Art. 83 GDPRGeneral conditions for imposing administrative fines. (n.d.). Retrieved from gdpr-info: https://gdpr-info.eu/art-83-gdpr/

Bhatia, P. (n.d.). *A summary of 10 key GDPR requirements*. Retrieved from advisera: https://advisera.com/articles/a-summary-of-10-key-gdpr-requirements/

Brodwall, J. (2019, 07 10). *Privacy and GDPR: What all developers should know - Johannes Brodwall*. Retrieved from Youtube: https://www.youtube.com/watch?v=6SHc7DWDDs4

Choudhary, A. (2023, 05 17). *Understanding CAP Theorem: Basics and Real-World Examples*. Retrieved from ashvinchoudhary.medium: https://ashvinchoudhary.medium.com/understanding-cap-theorem-real-world-examples-and-databases-d1ce0d807dca

General Data Protection Regulation. (2024, 05 03). Retrieved from en.wikipedia: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

Özkan, F. (2021, 11 19). *CAP Theorem and Distributed Database Management Systems*. Retrieved from medium: https://medium.com/cstech/the-cap-theorem-8b1b94b1586c#:~:text=CA%20(Consistency%20and%20Availability)&text=Most%20relational%20databases%20like%20MySQL,consistent%20and%20available%20(CA).

Redis . (n.d.). Retrieved from dbdb: https://dbdb.io/db/redis

Romero, A. G. (2018, 11 12). *High-availability resides in the Backing Services*. Retrieved from medium: https://romgerale.medium.com/high-availability-resides-in-the-backing-services-9b27149f904f

The CAP Theorem in DBMS. (2023, 04 02). Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/the-cap-theorem-in-dbms/

What personal data is considered sensitive? (n.d.). Retrieved from commission.europa.: https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en#:~:text=personal%20data%20revealing%20racial%20or,sex%20life%20or%20

Jacobs, L. (2024). Software Requirements Specification (Unpublished manuscript), FontysICT.

Jacobs, L. (2024). Security Design (Unpublished manuscript), FontysICT.