

JUMI SH

Alexander Petersen, Lucas Hahn, Sebastian Cywinski, Serdar Sahin, Nina Dannhäuser

Projektbetreuung seitens Fachhochschule durch: Jens Ehlers

Projektlaufzeit: 14.03. - 04.07.2024

Ausgangssituation:

Die Leitung der Abteilung GemIT des Justizministeriums SH (Gemeinsame IT der Justiz und Staatsanwaltschaften) erhält per Mail einen monatlichen Auszug zu Internet-Statistiken durch den IT-Dienstleister Dataport. Dieser Auszug enthält die Top 30 der aus dem Landesnetz am häufigsten aufgerufenen Webseiten sowie Top 30 der größten Download-Mengen einzelner Webseiten.

Es soll ein automatisierter Workflow entwickelt werden, über den die Daten aus dieser Internet-Statistik in ein automatisch erstelltes Reporting überführt werden sollen.

Folgende funktionale Anforderungen waren dafür gegeben:

- Die Eingangsdaten sollen in verarbeitbares Format umgewandelt werden
- Rand-Informationen aus der Mail sind zu entfernen und die bereinigten Daten sind in eine Datenbank zu überführen
- Es soll eine Webseite entwickelt werden zur Darstellung des Reportings
- Das System soll Möglichkeiten bieten, die Daten nach Zeiträumen zu filtern und nach Art der Werte (Anzahl Zugriffe oder Menge Download)
- Webseiten, die erstmalig im betrachteten Monat in den Daten vorkommen, sollen hervorgehoben werden
- Das System soll fähig sein, auf einem Webserver mit Anbindung an eine Datenbank zu laufen

Unsere Umsetzung:

Unser Team hat einen Javascript-Webserver eingerichtet. Dieser wird mittels HTML, JS und CSS programmiert. Serverseitig nutzen wir NodeJS mit Express als Framework. Der Server verbindet sich mit der lokalen Installation einer MySQL-Datenbank. Der Nutzer interagiert dann über einen beliebigen Browser mit dem System. Dort können die Datensätze importiert (**Extraction**), die grafisch aufgearbeiteten Daten betrachtet, verwaltet und auf verschiedene Weisen exportiert werden. Zunächst war es die Anforderung, die E-Mails als *txt*-Dateien importieren zu können. Um die Datensätze aus den E-Mails extrahieren zu können, wurde ein eigener Parser geschrieben (**Transformation**). Nach der Verarbeitung durch den Parser werden die Daten in die entsprechenden Tabellen in der Datenbank geschrieben (**Loading**).

Gegen Ende des Projektes fiel jedoch auf, dass seit Beginn des Jahres 2024 die Datensätze nicht mehr im selben Format vorliegen wie bisher. Stattdessen werden die Datensätze als *csv*-Dateien geliefert und können nach der Anpassung des Parsers ebenso importiert werden.

Zu Beginn sollte Docker genutzt werden, um die Lauffähigkeit auf verschiedenen Systemen sicherzustellen und alle Softwarekomponenten bündeln zu können. Jedoch wurden wir hier durch die strengen Sicherheitsvorgaben des Ministeriums beschränkt.

Für die interne Zusammenarbeit im Team wollten wir anfangs Github als Versionsverwaltung und Jira als Projektplanungstool nutzen. Später sind wir von Jira auf Github selber umgestiegen, um die Issues direkt mit den Commits verbinden zu können.

Während der Laufzeit des Projektes entstanden immer wieder neue Anforderungen und Verbesserungsvorschläge seitens des Kunden, die wir umgesetzt haben. Dazu zählen:

- Automatische Sicherheitsabfrage der Webseite durch Abfrage einer API (urlscan.io)
- Kommentarfeld für jede Webseite
- Webseiten in verschiedene Kategorien einteilen können (sicher/unsicher, wahrgenommen/nicht-wahrgenommen)
- Vergleich von Webseiten
- Import von csv-Dateien
- Light- / Dark Mode
- Exportfunktion als PDF/CSV
- Visualisierung von Datensätze zu einer Webseite
- Feedback an den Nutzer