Urbane Datenplattformen in der Cloud

Faruk Catal, Nikolay Tcholtchev, Philipp Lämmel (Frauhofer FOKUS) und Ina Schieferdecker (Fraunhofer FOKUS / TU Berln)

Table of Contents

Ziele urbaner Datenplattformen

Mit urbanen Datenplattformen verbinden viele Kommunen und Städte einerseits ein noch schwer zu fassendes Konzept und andererseits die Möglichkeit für mehr Innovation, Transparenz, Beteiligung als auch zusätzlichen Einnahmen für die Haushaltskasse. Dieser Artikel gibt einen Einblick zum Stand urbaner Datenplattformen und diskutiert das Pros und Cons direkter vs. indirekter Einnahmen aus öffentlichen Daten.

Mit urbanen Plattformen können Städte und Regionen wie Digital-Unternehmen an vorderer Front der Digitalisierung stehen. Verschiedene Nutzungsszenarien werden derzeit spezifiziert, diskutiert und pilotiert. Dabei handelt es sich im Allgemeinen um Szenarien, bei denen diverse Arten von Daten eine zentrale Rolle spielen. Im Zuge dessen werden die Daten an eine logisch zentralisierte, aber physisch verteilte Datenplattform – die sogenannte urbane Datenplattform – geleitet, die zudem verschiedenste Dienste zur Verwaltung, Kombination, Aufbereitung und Visualisierung der Daten für vielfältige Nutzungsszenarien bereithält. Dementsprechend werden smarte Städte und Regionen als datengetriebene IKT-basierte Ecosysteme verstanden, die ein Zusammenspiel verschiedenster Akteure über eine Reihe von stark verteilten Komponenten implementieren.

Dabei besteht das Ziel einer Urbanen Datenplattform darin, städtische Daten über eine logisch zentralisierte, physisch verteilte Infrastruktur anzubieten, damit eine Grundlage für die Entwicklung von Applikationen und Diensten entsteht, die einen entsprechenden Mehrwert im urbanen Raum erzeugen. Beispiele für relevante urbane Daten sind durch Kartenmaterial, Verkehrsinformationen, Wetterdaten, statistische Daten und allgemeine Verwaltungsdaten (z.B. Beschlüsse, Bevölkerungsdaten, Wahlergebnisse …) gegeben. Die Zusammenführung verschiedenster Datenquellen führt einerseits zur erhöhten Transparenz zwischen einzelnen Silos (z.B. Abteilungen innerhalb einer Behörde) und ermöglicht andererseits die effiziente Überwachung und Kontrolle der städtischen Prozesse, sowie die Gewinnung von Erkenntnissen (wie im folgenden erläutert) und entsprechende (strategische) Reaktionen auf städtischer Ebene. Zusätzlich können die Daten von Datenjournalisten genutzt werden, sowie

zu einer verbesserten Transparenz der städtischen und politischen Entscheidungsprozesse beitragen – z.B. offene statistische Daten können als Argumentationsgrundlage bei der Begründung bestimmter Entscheidungen verwendet werden. Darüber hinaus wird die Nutzung von urbanen Daten in (mobilen) Applikationen angestrebt – Beispiele für solche Applikationen/Dienste sind durch Navigationssysteme, touritische Applikationen, Informationsportale und Frühwarn-Apps/Dienste gegeben. Konzeptionell besteht auch die Möglichkeit für eine Bürgerbeteiligung. Zum Beispiel existieren Dienste, die das Melden von Strassenschäden ermöglichen und dadurch die Teilnahme von Bürgern und die Partizipation bei Instandshaltungsentscheidungen unterstützen.

Urbane Daten können aus verschiedensten Quellen, beispielsweise aus öffentlichen, industriellen oder privaten Quellen, stammen. Die Bereitstellung von urbanen Daten wird durch anhaltende Initiativen beim Öffnen von behördlichen Daten unterstützt, geht jedoch noch darüber hinaus. Mit Hilfe von urbanen Daten können fundierte Entscheidungen im alltäglichen administrativen Geschäft getroffen werden. Das PAS Smart City Konzeptmodel (BSI, 2016) definiert vier Typen an Erkenntnissen, die aus urbanen Daten gewonnen werden können:

1. **Betriebliche Erkenntnisse**, um Eigenschaften und Charakteristiken von urbanen Sachverhalten und Prozessen zu verstehen, um dadurch Verbesserungsoptionen ableiten zu können - z.B. Verkehrsinformationen in Echtzeit um ein verbessertes multi-modales Routing zu erreichen.
2. **Kritische Erkenntnisse**, um aktuelle Vorfälle zu beobachten und dadurch Handlungsempfehlungen ableiten zu können – z.B. Transprarenz und Hinterfragen von politischen Entscheidungen, Melden von Strassenschäden und verwandte Formen der Bürgerpartizipation.
3. **Analytische Erkenntnisse**, um Muster und Korrelationen zu identifizieren und dadurch Vorbedingungen für urbane Innovation, Auswirkungseinschätzungen oder Herausforderungen und Möglichkeiten bei der urbanen Entwicklung ableiten zu können – z.B. Verschnitt von touristischen und Gastranomie relevanten Daten, oder von statistischen Mobilitätsdaten und Infrastrukturinformationen inklusive Zustandsinformationen wie Strassenschäden.
4. **Strategische Erkenntnisse**, um einen allumfassenden Ansatz bei den strategischen Zielen, Plänen und Entscheidungen innerhalb der urbanen Umwelt zu ermöglichen – z.B. Einfluss auf längerfristige politische Entscheidungen wie Mietbremse auf Basis des Verschnitts verschiedener Statistiken.

Rahmenbedingungen für urbane Datenplattformen

Was vor zehn Jahren als Bewegung zur Bereitstellung und Öffnung von Daten der öffentlichen Hand begann (siehe beispielsweise (Sunlight Foundation, 2007) oder (Both & Schieferdecker, 2011)), hat sich mehr und mehr dahin weiterentwickelt, dass nicht nur Daten der öffentlichen Verwaltung – wie in Berlin (Berlin Online & FOKUS, 2011), Deutschland (Seitenbau & FOKUS, 2013) oder Europa (Cap Gemini et al, 2015) – sondern auch Daten anderer Akteure und Branchen bereitgestellt werden. So wird die Entwicklung von Open Science-Datenplattformen wie Open Power System Data (Neon, 2015) oder Transparenz-Datenplattformen wie Netzdaten-Berlin (Stromnetz & FOKUS, 2012) gefördert.

Darüber hinaus wächst das Verständnis, dass ein ganzes Datenuniversum zu etablieren ist: angefangen bei hochkritischen sicherheitsrelevanten Daten über ebenso zu schützende personenbezogene, kommerzielle und das Gemeinwohl betreffende Daten, bis hin zu vollkommen offenen Daten (entsprechend Sunlight Foundation, 2007). So rücken Daten und Informationen in zunehmendem Maße als Gut bzw. Ressource in Datenökonomien in das Zentrum der Betrachtung, die auch für Kommunen und Städte ihre Wirkmächtigkeit entfalten können.

Die Europäische Kommission bereitet mit der „Digitalen Agenda Europas“ (EU, 2010) den EU-Binnenmarkt auf das digitale Zeitalter vor. Im Mittelpunkt stehen Daten als Basisressource der neuen wissensbasierten Gesellschaft und Wirtschaft sowie, darauf aufbauend, neue branchenübergreifende Geschäftsmodelle und innovative Dienste. Daten und datengetriebene Dienste verschaffen einen Überblick, verhelfen zu neuen Schlussfolgerungen, bringen mehr Informiertheit und Rationalität in politische und wirtschaftliche Diskurse. Im Kontext der „European Innovation Partnership Smart Cities and Communities“ wurde unter anderem eine von Städten, Kommunen, Wirtschaft und Wissenschaft getragene Initiative zur Etablierung einer interoperablen Plattform-Architektur für Wirtschaften, Arbeiten und Leben im urbanen Raum etabliert (EIP SCC, 2015), so dass öffentliche Verwaltungen urbane Plattformen kostengünstig umsetzen und betreiben und darüber Innovationen, Beteiligung und Transparenz befördern können. Auch wenn sich die Initiative vor dem Hintergrund der exponentiell wachsenden Urbanisierung auf Städte konzentriert, so sind viele der Ansätze, Konzepte und Technologien auch auf Kommunen und den ländlichen Raum übertragbar. Einen Eindruck von den Elementen einer solchen urbanen Plattform vermittelt Abb. 1. Für das Aufsetzen und den Betrieb einer solchen Plattform müssen die Details von Technikern verstanden und beherrscht werden. Die Verwaltung muss darüber hinaus ein Verständnis für die Fähigkeiten, Möglichkeiten und Varianten einer urbanen Plattform entwickeln, um sie entsprechend der Ziele und Prioritäten nutzen zu können. Eine Grundidee interoperabler urbaner Plattformen ist es, dass sie aus Komponenten verschiedener Hersteller bestehen und so eine inflexible Kopplung an einzelne Anbieter vermieden wird. Dafür müssen die Schnittstellen und Formate der urbanen Plattform offen sein und den Vorgaben der Referenzarchitektur folgen. Zudem sind ausgewählte Komponenten auch als Open Source (beispielsweise für Open Data oder Open IoT verfügbar.

Auch in Deutschland formieren sich Initiativen zur Bereitstellung und Nutzung von urbanen Daten über derartige Plattformen, wie beispielsweise die Standardisierungsinitiative beim DIN (DIN, 2017) zeigt. Neben den technischen Fragestellungen sind administrative und ggfs. politische Hindernisse zu überwinden. Beispielsweise liegen Daten verteilt auf Endgeräten und Servern – zentralisierte Ansätze zur Bereitstellung der Daten verbieten sich auch aufgrund von technischen Fragen wie Skalierbarkeit. Zudem stellt der Verlust insbesondere für

Unternehmen und Institutionen ein großes Risiko dar, so dass die Daten geeignet zu schützen als auch vor Veraltung oder Modifikation zu bewahren sind. Ebenso ist die Datensicherheit entsprechend der Eigentümer-/Nutzerverhältnisse und rechtlicher Bestimmungen zu gewährleisten.

Zur Bereitstellung sind die Daten zu kategorisieren, um sie so den geheimen (einer beschränkten Nutzergruppe zugänglichen), kommerziellen (den Kunden zugänglichen) oder offenen Daten (allen zugänglichen) zuzuordnen. Zudem sind sie nach Herkunft, inhaltlichem Bezug, technischen, qualitätsorientierten und rechtlichen Eigenschaften, und ihrer Verfügbarkeit inkl. Zugriffsmöglichkeiten, Form und Kosten zu charakterisieren. Wesentlich ist ebenso, dass urbane Daten in sogenannten maschinen-verarbeitbaren, wohl-definierten und gut dokumentierten Formaten und stabilen Nutzungsbestimmungen, d. h. unter bekannten und zuverlässigen Bereitstellungs-, Aktualisierungs- und Korrekturraten zur Verfügung gestellt werden. Nur bei digitaler Verfügbarkeit, rechtlich und technisch wohldefinierten Nutzungsbestimmungen und zuverlässiger Bereitstellung lassen sich die in vielfältigen Studien prognostizierten Mehrwerte urbaner Daten realisieren (siehe (TSB, 2014) und (DIVSI, 2016)).

Dabei ist der Nutzen urbaner Plattformen für die jeweilige Zielgruppe zu definieren. So stehen z. B. für Bürger, Touristen und sonstige Privatkunden Informationen über kommunale Angebote, Bürgerbelange, Bürgerdienste, etc. im Vordergrund. Für Unternehmen liegt der Nutzen eher in der Generierung neuer Informationsprodukte und -dienste durch die Bereitstellung eigener Daten an Andere und durch die Nutzung der Daten Anderer.

Zudem sind neben den allgemeinen Anforderungen an Aktualität, Qualität, Verfügbarkeit der Daten, endgeräteübergreifendem Zugang und rechtsicherer Nutzung auch die IT-Sicherheit und der Datenschutz zu berücksichtigen. Grundsätzlich stehen urbane Daten und ihre teilweise Offenheit in einem allgemeinen politischen Spannungsfeld, aber ebenso konkret zu Werten des Persönlichkeits- und Datenschutzes. Dazu sind unterschiedliche Schutzkategorien von Daten sowie geeignete Identity-Management-Mechanismen und Zugriffsmechanismen zu etablieren, die eine effektive Steuerung der Datennutzung ermöglichen. Zudem ist eine mögliche Manipulation von Daten auszuschließen. Insbesondere für Unternehmen, die auf der Basis urbaner Daten neue Informationsprodukte und -dienste entwerfen und anbieten, ist die Integrität der Daten eine kritische Geschäftsgrundlage.

Aufbauend auf den urbanen Daten können in nächsten Schritten neue datengetriebene Angebote beispielsweise für Handel, Logistik, Verwaltung, öffentliche Sicherheit und Kommunikation entworfen werden. So werden oftmals Anwendungen aus den folgenden Bereichen umgesetzt:

- Vernetzte, multi-modale Verkehrsführung
- Vernetzte urbane Infrastrukturen wie intelligente Straßenbeleuchtung
- Vernetztes Notfallmanagement
- Digitalisierte Verwaltung und Beteiligung
- Dezentrale Energieversorgung

Durch die Verschneidung von verschiedenen urbane Daten kann so ein essentieller Mehrwert generiert werden. So können urbane Daten eine Grundlage für fundierte Entscheidungen im alltäglichen Geschäft von Unternehmen als auh den administrativen Abläufen der Verwaltung sein. So können Kommunen durch die Bereitstellung urbaner Daten über die daraus resultierenden neuen Dienste, beispielweise zur Unterstützung von Entscheidungsprozessen bei der Stadtentwicklung oder Verkehrsplanung, profitieren.

Dabei ist, wie wir bereits in (Klessmann et al, 2012) formuliert haben, die geldleistungsfreie Bereitstellung öffentlicher Daten den geldleistungspflichtigen Ansätzen vorzuziehen. Die geldleistungsfreie Datennutzung folgt dem Open-Data-Gedanken, erhöht die Datennutzung und fördert die volkswirtschaftliche Wertschöpfung (siehe auch wie bereits erwähnt (TSB, 2014)). Die nichtkommerzielle Nutzung öffentlicher Daten sollte grundsätzlich geldleistungsfrei sein. Eine geldleistungsfreie Bereitstellung minimiert gleichsam die Verwaltungs- und Abrechnungsaufwände in der öffentlichen Hand.

Jedoch sind öffentliche Daten wie oben beschrieben nur ein Teil urbaner Daten. Es gibt eine Vielzahl von urbanen Daten mit hoher Attraktivität und ökonomischen Potenzialen. Deren aktuelle, feingranulare und hochqualitative Aufbereitung ist andererseits kostenintensiv, so dass die Kosten den ökonomischen Potentialen gegenüber gestellt werden sollten.

Vor dem Hintergrund der Heterogenität bei den Geldleistungsmodellen und zur Förderung einer ebenenübergreifenden Kompatibilität sollten gemeinsame Grundsätze zur Bepreisung der Datennutzung in Deutschland, ggfs. in Europa vereinbart werden. Dazu sollten einige Eckpunkte berücksichtig werden:

- Die Bepreisung der bereitgestellten Daten orientiert sich am Zweck ihrer Nutzung. Geldleistungen sollten nur auf Dienste mit Mehrwertcharakter und Daten mit hohem Pflegeaufwand erhoben werden.
- Die Erhebung von Geldleistungen für die Bereitstellung und Reproduktion von Daten für Dritte muss wirtschaftlich erfolgen und durch Zusatzaufwand gerechtfertigt sein, z.B. durch die regelmäßige Aktualisierung von großen Datenmengen.
- Die Bemessungsgrundlage für die Kalkulation von Geldleistungen ist auf die ermittelten Zusatzkosten für die Bereitstellung und Reproduktion von Daten für Dritte zu beschränken (Kostendeckung). Die Höhe der in Rechnung zu stellenden Zusatzkosten ist nach betriebswirtschaftlichen Methoden zur Preiskalkulation zu ermitteln.
- Die Ermittlung von Geldleistungen sollte für die Verwaltung einfach und für den Nutzer nachvollziehbar sein. Die Anzahl der Parameter zur Ermittlung der Geldleistung sollte minimal sein. Ein gemeinsames Kalkulationsschema für die Bepreisung von Daten sollte zur Orientierung potenzieller Datenbereitsteller entwickelt werden.

Die Erhebung von Geldleistungen muss wirtschaftlich erfolgen. Sollte eine Erhebung von Geldleistungen durch administrativen Aufwand, wie Rechnungsstellung, Zahlungsverfolgung, Rechnungswesen etc., für die öffentliche Verwaltung unwirtschaftlich sein, ist von der Geldleistungspflicht Abstand zu nehmen.

Unsere Lösungen

Unsere Lösungen haben wir über eine Vielzahl von Datenplattformen und -portalen schrittweise entwickelt und ausgebaut. Die Abbildung 1 verdeutlicht die Architektur der GovData-Plattform[1], die die Realisierung des deutschen Open Government Data-Portal ist. Die

[1] Die GovData-Plattform (https://www.govdata.de) stellt offene Verwaltungsdaten bereit. Der Pilot dieser Plattform wurde 2012 von Fraunhofer FOKUS unter der Leitung des Bundesministeriums des Inneren entwickelt, Februar 2013 begann der erfolgreiche Pilotbetrieb und Anfang 2015 der Produktivbetrieb.

GovData-Plattform wurde aufbauend auf unserem Berliner Datenportal daten.berlin.de und dem Stromnetz-Datenportal netzdaten-berlin.de entwickelt. Zur Nachnutzung durch andere Kommunen wurde sie zudem konsequent als offene Software entwickelt und wird über GitHub bereitgestellt. So ermöglichen wir eine freie Nutzung und Betrieb der Plattform und beugen einem Hersteller Lock-in vor. Die Basis für die Entwicklung der GovData-Plattform bietet die Portalsoftware *Liferay* (Liferay, 2017), die genau wie die Plattform selbst in Java entwickelt wurde. Für den (Pilot- und später Produktiv-)Betrieb wird das webbasierte System *Comprehensive Knowledge Archive Network (CKAN)* (OKFN, 2017) zum Speichern und Verteilen von Metadaten verwendet. Als Datenbank kommt *PostgreSQL* (PostgreSQL, 2017) zum Einsatz.

Eines der vorrangigen Ziele der GovData-Plattform ist die einfache Bereitstellung von Daten und der einfache Zugang zu diesen, beispielsweise über innovative Applikationen und Dienste (z.B. Navigationssysteme und mobile Applikationen, wie sie auf der rechten Seite der Abbildung 1 dargestellt sind). Die GovData-Plattform bietet zudem den Nutzern die Gelegenheit, einen Einblick in öffentliche Daten zu kriegen und Daten über eingebaute Such- und Filterfunktionen zu finden, selektieren, analysieren und auszuwerten. Entwicklern wird über eine entsprechende Application Programming Interface (API) der Zugriff auf die gespeicherten Metadaten gewährt. Auf diesem Wege können Entwickler aus der Community eigene Anwendungen unter Verwendung der offenen Daten entwickeln.

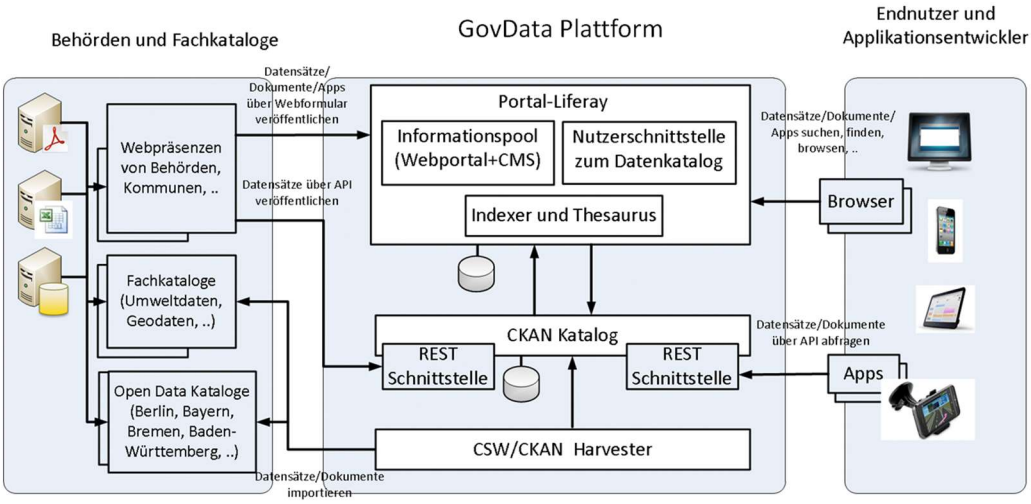

Abbildung 1: Architektur der GovData-Plattform

So gibt es diverse Ansätze, die die praktische Anwendung von Open Data-Plattformen behandeln. In (Tcholtchev, et al., 2012) diskutieren Tcholtchev et. al. über die Umsetzung von Open Data-Strategien für die Realisierung eines Mobilitätskonzepts. Bei diesem Beispiel werden Elektrofahrzeuge über eine in einer Cloud-Umgebung laufenden Open Data Plattform mit Daten versorgt und können Informationen über aktuellen Ladezustand, Position oder ähnliches der Plattform mitteilen.

Dabei ist zu beachten, dass Systeme, die Daten mit einer bestimmten Effizienz bearbeiten und ausfallsicher arbeiten müssen, hohe Rechnerkapazitäten erfordern. Einfache Server können bei Lastspitzen nicht immer die geforderte Leistung erbringen. Der Einsatz von Cloud-Computing bietet Lösungsmöglichkeiten, die im Weiteren beschrieben und analysiert werden. Sowohl die Herangehensweise an eine Migration einer solchen Plattform mit Hilfe verschiedener Virtualisierungslösungen in die Cloud, als auch die Vorteile des Betriebes dieser Software in einer Cloud-Umgebung werden präsentiert.

Es ist keine Seltenheit, dass Dienste oder Server bei Spitzenzeiten nicht jede Anfrage eines Klienten erfolgreich bedienen können. So geschah es auch bei GovData. Neben dem einfachen Betrieb ist die Möglichkeit der flexiblen Skalierung von Ressourcen ein wesentlicher Grund, der das Wachstum von Cloud-Diensten vorantreibt und den Entschluss vieler Unternehmen zur Auslagerung der eigenen IT-Infrastruktur in die Cloud begünstigt. Unter anderem beschäftigen sich internationale und nationale Cyber-Sicherheitsagenturen (z.B. die ENISA (ENISA, 2017) auf EU-Ebene) damit, die Rahmenbedingungen für sichere und resiliente Cloud-Deployments auszuarbeiten.

Die Entwicklung der Cloud-Variante für GovData erfolgte unter Verwendung von Open Source-Software. Während die Enterprise-Edition der eingesetzten Basissoftware (Liferay (Liferay, 2017)) einige Methoden zur Lastenverteilung und Clustering mitbringt, muss man diese bei der verwendeten Community-Edition manuell umsetzen. Um auch den Vorgang der Migration in die Cloud und den späteren Betrieb der Plattform in einer Cloud-Umgebung weiterhin mit minimaler Inanspruchnahme kommerzieller Techniken zu verwirklichen, werden die Schwerpunkte dieses Artikels vor allem auf der Verwendung von Open Source-Software liegen.

Realisierung der Migration der GovData-Plattform in die Cloud

Eine neu zu entwickelnde Applikation, die in einer Cloud-Umgebung betrieben werden soll, kann bereits so geplant und implementiert werden, dass sie den Ansprüchen der Cloud-Umgebung genügt. In dem Fall wäre die Frage nach der Migration hinfällig. Meist wird über den Einsatz der Cloud-Technik in den Betriebsablauf erst im späteren Prozess der Entwicklung entschieden. Hier lässt sich die Anpassung oder Optimierung der Applikation für die Cloud-Umgebung nicht vermeiden. Es gibt viele Standardisierungsversuche von verschiedenen Firmen für den Vorgang einer Migration in die Cloud, jedoch keine allgemein anwendbare Migrationsstrategie. Je nach Struktur der Applikation, der verwendeten Entwicklungsumgebung, der Programmiersprache und weiteren Faktoren ändert sich der Ablauf einer Migration.

Im Folgenden wird die virtualisierte Architektur der GovData-Plattform erläutert. Die Abbildung 2 zeigt eine Übersicht der eingesetzten VM-Komponenten, welche nachfolgend kurz beschrieben werden.

1. Apache Load Balancer (LB): Dieser nimmt die Client-Anfragen entgegen und leitet diese zu einer geeigneten Liferay-VM weiter.
2. Dies sind die erzeugten Liferay-VMs, die die vom LB weitergeleiteten Anfragen bedienen.
3. Der CKAN-Rechner (3) hat direkten Zugriff auf einen Datenbankrechner (4).
4. In dieser VM (4) läuft der PostgreSQL Datenbankserver, der die Datensätze von CKAN verwaltet.
5. (5) ist die PostgreSQL Datenbank VM, in der die Daten von Liferay verwaltet werden.
6. Dieses Symbol (6) stellt eine Ordnerfreigabe dar, in der einige Daten für die einzelnen Liferay-VMs bereitgestellt werden, die für die Liferay Oberfläche benötigt werden.

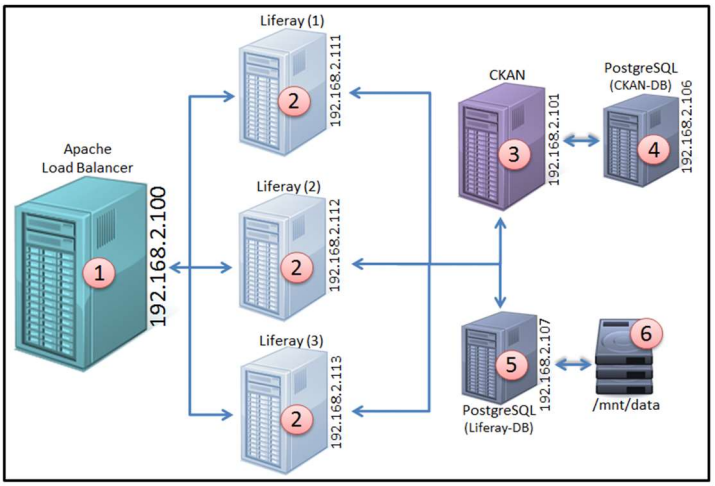

Abbildung 2: Struktur der VM-Komponenten

Um die oben abgebildete Architektur praktisch umzusetzen, ist Vorarbeit erforderlich. Bevor die virtuellen Maschinen für den Betrieb konfiguriert werden, müssen diese mit dem passenden Betriebssystem bespielt werden, wobei in dieser Fallstudie die Ubuntu Server Edition in der Version 12.10 zum Einsatz kam. Initial werden virtuelle Maschinen mit einer CPU und zwei GB Arbeitsspeicher verwendet und die entsprechende Software-Komponenten auf die VMs installiert. Sind die fünf VMs fertig eingerichtet und in die *vSphere* Umgebung beziehungsweise in den *VirtualBox VM-Host* übertragen worden, kann nach anschließender Einrichtung des Netzwerks ein erster Betriebstestlauf durchgeführt werden. Hier können bereits Fehler entdeckt werden, die zum Beispiel auf eine falsch konfigurierte Datenbank hinweisen könnten. Wie in der Abbildung 2 erkennbar, wurden für die benötigten zwei Datenbanken auch zwei separate VMs verwendet. Eine einzelne leistungsfähige VM mit zwei Datenbanken, die jeweils von CKAN und Liferay genutzt werden, wäre auch möglich. Zur Ermittlung der Tragfähigkeit einer aktiven Liferay-VM wird ein erster Lasttest ausgeführt, bei dem ermittelt wird, wie viele Anfragen eine einzelne Liferay-VM bedienen kann, ohne dass dabei Fehler entstehen. Diese Information ist für die Konfiguration der entwickelten Steuerung des VM-Server notwendig (siehe Abbildung 3). In der Abbildung 3 wird die Funktionsweise der Steueranwendung abgebildet. Mit steigender Anzahl der Klienten – und damit der Anfragen an die GovData-Plattform – werden weitere virtuelle Maschinen erzeugt und aktiviert. Damit stehen mehr Ressourcen zur Verfügung und somit können mehr Anfragen ohne Fehler bedient werden. Entsprechend diesem Vorgang sollen die nicht benötigten VMs beim Wegfall der Last auch wieder deaktiviert werden. Dieser Prozess entspricht der Elastizität einer Cloud auf ”Infrastructure-as-a-Service (IaaS)” Ebene.

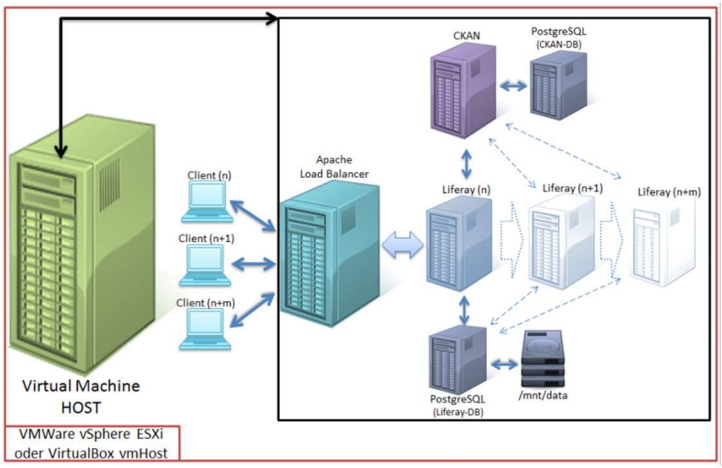

Abbildung 3: Host für die virtuellen Maschinen

Die Lastabfrage der einzelnen Maschinen wurde über den Apache Load Balancer realisiert. Es gibt auch hier zwei Lösungsansätze. Bei dem ersten wird davon ausgegangen, dass die maximale Anzahl an erzeugbaren VMs bereits bekannt ist. Dies erleichtert die Arbeit bei der Konfiguration des Load Balancers. Hierdurch kann die vordefinierte Anzahl an Liferay-VMs im Voraus in der Balancer Konfiguration berücksichtigt werden und erfordert keinen Neustart des Apache Servers nach Veränderung der Konfigurationsdateien. Die Lasttests, die vor Inbetriebnahme der VM-Steuerung gemacht wurden, lassen einen Rückschluss zu, wie viel Liferay-VMs man maximal für einen stabilen Betrieb benötigen würde. Der zweite Lösungsweg setzt keine obere Grenze für die Anzahl der Liferay-VMs und erfordert auch keine vorherige Definition dieser Größe. Das verlangt jedoch einen Neustart des Apache Servers nach einer Veränderung der Konfiguration. Der Neustart des Apache Servers dauert in der Regel nur wenige Sekunden und wird automatisch vorgenommen. In dieser kurzen Zeit wäre die Anwendung jedoch nicht erreichbar. Als eine Weiterentwicklung der zweiten Vorgehensweise wäre die stufenweise Erhöhung der Anzahl der VMs, beispielsweise könnten bei jedem Neustart des Apache Load Balancers zehn zusätzliche virtuelle Maschinen in der Apache-Konfiguration berücksichtigt werden. Damit würde man das Neustarten des Apache Servers minimieren und hätte dennoch den Vorteil der stetigen Erweiterung der virtuellen Maschinen.

Server- oder Datenbank-Clustering ist ein Schritt auf dem Weg zur Hochverfügbarkeit von Webdiensten. Ein weiterer möglicher Umgang mit der Problematik der Überlastung von Servern ist das proaktive Verteilen der Last, noch vor dem Eintreffen, auf den geeigneten Server. Hier bietet Apache einen leicht zu konfigurierenden *Load Balancer*, der verschiedene Methoden des Lastverteilens unterstützt. Load Balancing ist ein wichtiger Aspekt der Hochverfügbarkeit. Ein einzelner überlasteter Server wird somit als Fehlerquelle ausgeblendet. Ist der *Apache Load Balancer* richtig eingerichtet, ist eine wichtige Grundlage zur Hochverfügbarkeit von Liferay geschaffen. Die Konfigurationsarbeit ist jedoch damit nicht zu Ende. Vielmehr beginnt die Konfigurationsarbeit erst hinter dem Load Balancer, weil hier eine Abstimmung der einzelnen Komponenten aufeinander zu bewerkstelligen ist. Die größte Hürde ist die Synchronisierung der einzelnen Liferay Instanzen. In dem Fall von GovData wird diese Synchronisierung über eine gemeinsame, hoch performante Datenbank und über ein gemeinsames Dateisystemcluster erreicht.

Bei der Beschreibung der Liferay-Hochverfügbarkeit werden Kombinationen aus Hardware und Software basierenden Load Balancern vorgestellt. Dabei existieren wiederum Varianten, die mit einem Datenbank-Cluster auskommen und andere, die mehrere Datenbanken benutzen. Der Einsatz dieser Methoden ist jedoch in der Enterprise-Edition vorgesehen. In der

Community-Edition, die für die GovData-Plattform genutzt wurde, ist man auf eigene Mittel zur Umsetzung der Liferay Hochverfügbarkeit angewiesen. Es müssen verschiedene Anwendungsebenen beachtet werden, die eine Abstimmung voraussetzen. Die durchgeführten Tests haben gezeigt, dass auch nach vielen aktivierten, gemeinsam arbeitenden Liferay-Knoten, eine einzelne hoch performante Datenbank keinen Leistungsverlust hervorruft. Daher wurde bei der technischen Umsetzung nur eine gemeinsam genutzte Datenbank eingesetzt, die in einer virtuellen Maschine betrieben wurde.

Die Ergebnisse der cloud-basierten GovData-Version

Um die Last- und Performance der migrierten Lösung zu testen, wurden diverse Tests mit Hilfe des *JMeter* Tools (JMeter, 2017) (Ramakrishnan, Shrawan, & Singh, 2017) durchgeführt. JMeter bietet eine breite Palette an Testmöglichkeiten an. Bei den folgenden Lasttests wurden einfache HTTP-Anfragen an den Apache Load Balancer geschickt und die Antworten des Servers untersucht. Es wurden dabei drei Anfragen zusammengefasst:

1. Anfrage auf die Hauptseite der Webapplikation (1),
2. Anfrage auf eine Unterseite der Webapplikation (2) und
3. Anfrage auf einen frei gewählten Datensatz aus der Datenbank (3).

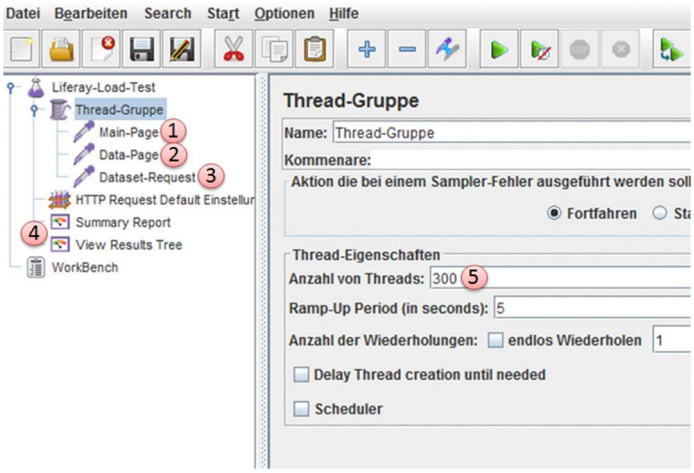

Abbildung 4: JMeter Benutzeroberfläche zur Test-Konfiguration

JMeter beinhaltet eigene Analysewerkzeuge (siehe (4) in Abbildung 6), wodurch der Einsatz weiterer Software zur Auswertung der Ergebnisse entfällt. Bei der Durchführung der HTTP-Anfragen wurde der erste Durchgang mit 300 (siehe (5) in Abbildung 6) gleichzeitigen Threads begonnen. Diese wurden bei jedem Durchgang um 100 Threads erhöht und jeder einzelne Durchgang wurde mindestens dreimal wiederholt, um einzelne Messfehler auszuschließen. Für den ersten Durchlauf wurden 300 Threads ausgewählt, weil vorherige Tests gezeigt haben, dass ab diesem Wert messbare Fehler entstehen. Die bei den Tests verwendeten Liferay-VMs besaßen 4-GB Arbeitsspeicher. Um jeden Testdurchlauf unter gleichen Voraussetzungen durchzuführen, wurden nach dem Abschluss jeder Stufe die Liferay VMs heruntergefahren und neu gestartet.

Die Abbildung 7 zeigt die Funktionsweise von JMeter. Nach erfolgreicher Konfiguration eines Testlaufs, kann ein Zielserver mit vielen Aufrufen angefragt werden, um somit die Grenzen der Belastbarkeit zu überprüfen. Dieser Vorgang erfordert auf der ausführenden Seite einen Rechner mit genügend hoher Leistung. Andere Prozesse sollten während der Ausführung der Tests nicht gestartet werden, da diese die Testergebnisse verfälschen können.

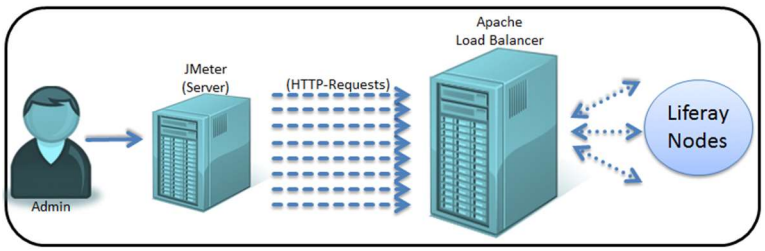

Abbildung 5: JMeter einfaches Testen

Die Praxis hat gezeigt, dass nach der Einrichtung des Load Balancers eine sehr hohe Anzahl an gleichzeitigen Threads benötigt werden, um den zu testenden Server an dessen Grenzen zu befördern. Die Abbildung 8 gibt für solch ein mögliches Problem eine Lösung. JMeter unterstützt unter anderem die Option des verteilten Testens. In diesem Fall werden einzelne verfügbare Rechner oder VMs mit der Client-Version von JMeter eingerichtet. Diese werden über den JMeter Server mit den Testfällen automatisch versorgt und zur Ausführung der Tests ferngesteuert. Setzt man eine relativ hohe Anzahl von Client-Rechnern ein, können Server unter einer extrem hohen Last getestet werden. Solch ein Testvorgang kann für spätere Arbeiten ausgebaut und in einer Cloud-Umgebung als sogenannte Network-Testing-as-a-Service (NTaaS) Variante des Cloud-Computings betrieben werden.

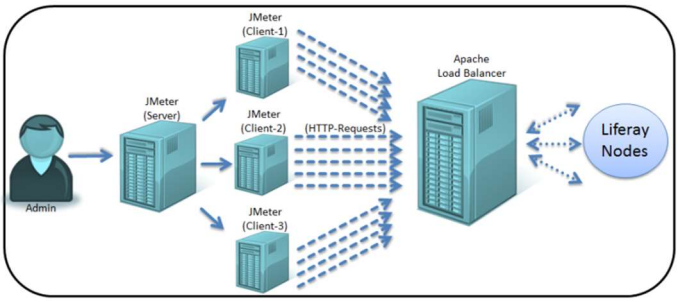

Abbildung 6: JMeter verteiltes Testen

Im Folgenden werden die Messungen der Performance der GovData-Plattform nach der Cloud-Migration analysiert. Abbildung 9 veranschaulicht diese Performance anhand der durchgeführten Experimente. Die der Auswertung zugrundeliegenden Messergebnisse zeigten, dass ohne eine tiefer gehende Optimierung der GovData-Plattform für eine sehr hohe Anzahl von Anfragen, diese ohne eine Lastenverteilung ca. 300 gleichzeitige HTTP-Anfragen mit je zwei Seitenklicks und einer Datensatzabfrage erfolgreich bedienen konnte. Der bei den durchgeführten Tests verwendete Zustand der GovData-Plattform entsprach nicht der Betriebsversion, bei der weitere Optimierungen bezüglich der Belastbarkeit vorgenommen wurden. Somit sind bessere Ergebnisse mit der endgültigen Version der Plattform zu erwarten. Die Ergebnisse verdeutlichen zwar bereits eine höhere Performance beim Einbeziehen von zwei virtuellen Maschinen, jedoch ist die Annahme, dass zwei einbezogene VMs, im Vergleich zu einer, die doppelte Leistung mit sich bringen, mit den durchgeführten Versuchen widerlegt worden.

Die Ergebnisse der durchgeführten Messungen mit verschiedenen Belastungen und verschiedener Anzahl an virtuellen Maschinen, auf die die Last über einen Load Balancer verteilt wurde, werden in der Abbildung 9 zusammengefasst dargestellt. Das spätere Eintreten der ersten Fehler beim Erhöhen der Anzahl der aktiven VMs auf zwei und die deutlich höhere Anzahl der gemessenen Fehler bei nur einer VM sind in der Abbildung gut ablesbar.

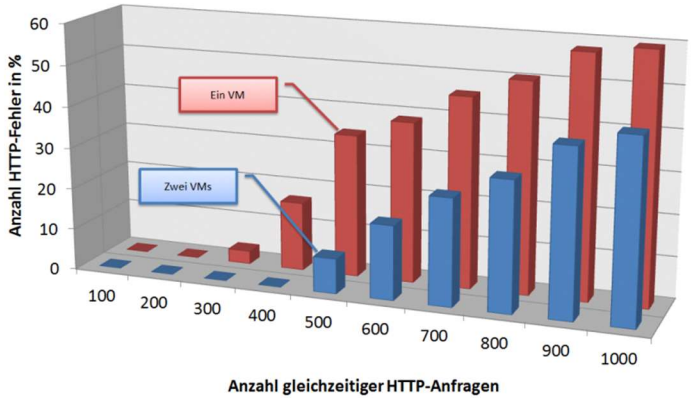

Abbildung 7: Graph über die Fehlerverteilung bei Einbeziehen von einer und zwei VMs

Herausforderungen in der Umsetzung

Da die meisten Verwaltungsmitarbeiter*innen als auch Politiker*innen keine Datenexpert*inn*en sind und ihnen die Erfahrung mit der Datenanalyse fehlt, ist es sinnvoll, auf Datenspezialisten zur Datenaufbereitung und Datenmanagement, Datenzugriff und Datenanalyse zurückzugreifen. Verschiedene Daten- und Metadatenportale wurden etabliert und sind heute im Einsatz. Im Rahmen dieses Artikels wird auf die von Fraunhofer FOKUS initial entwickelte GovData-Plattform, eine Sonderform einer urbanen Datenplattform, eingegangen. Über die GovData-Plattform soll mit Hilfe des Konzeptes von *offenen Daten* Organisationen, Unternehmen und Bürgern einer modernen Stadt der Zugriff auf öffentliche Daten zur gemeinsamen Gestaltung der städtischen Prozesse ermöglicht werden. Über eine solche Plattform soll darüber hinaus die aktive Bereitstellung und Nutzung der Daten zugänglich gemacht werden. Die folgende Definition fasst die Idee hinter Open Data zusammen (Definition, 2017):

> *"A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike."* (Definition, 2017)

Es ist zudem hervorzuheben, dass die GovData-Plattform, wie in Abbildung 1 dargestellt, vor allem die Verwaltung von Metadaten, das heißt Daten, die die eigentlichen Daten beschreiben, ermöglicht. Die Daten liegen dabei dezentral bei den Datenbereitstellern. Dabei wird die Metadaten-Struktur für Open Government Data in Deutschland genutzt. Die Ausweitung auf die Datensätze an sich und deren logisch zentralisierten Speicherung ist möglich und wurde im folgenden Papier (Scholz, Nikolay, Lämmel, & Schieferdecker, 2017) untersucht.

Zusammenfassung

Die Chancen und Herausforderungen an urbane Plattformen am Beispiel von GovData und der damit verbunden Bedarf an die Hochverfügbarkeit von GovData bildete die Ausgangslage dieses Artikels, wobei die Migration der Plattform in die Cloud mit Hilfe von Open Source-Werkzeugen im Vordergrund stand. Die Motivation der Migration ergab sich durch die erwarteten Vorzüge des Cloud-Computings: Erstens ist die Hochverfügbarkeit der GovData-Plattform im praktischen Betrieb notwendig und zweitens soll die dynamisch-elastische Reaktion der GovData-Plattform auf Lastspitzen gewährleistet werden.

Die Ansätze dieses Artikels greifen dabei Techniken des Cloud-Computings auf und untersuchen die Anwendung dieser Techniken in Kombination mit eigenen Erweiterungen, auf die GovData-Plattform. Die daraus resultierenden Technologien und Besonderheiten in

Bezug auf die GovData-Plattform wurden im Rahmen dieses Artikels identifiziert. Die präsentierte Umsetzung baut auf die Grundlagen des Cloud-Computings auf, dessen Wurzeln in der Virtualisierung von Rechenleistungen liegen.

Bei der Untersuchung von diversen Cloud Service Providern wie *Amazon* (Amazon, 2017) und *RackSpace* (RackSpace, 2017), aber auch von Open Data Cloud-Plattform-Anbietern wie *Junar* (Junar, 2017), stellt man schnell fest, dass sich ganze Unternehmungen auf der Basis vom geschickten Verpacken und Bereitstellen von bekannten Techniken, mit der richtigen Kombination von Softwarelösungen, auf dem Markt etabliert haben. Im Detail steckt dabei selbstverständlich jede Menge Knowhow und Forschung dahinter. In dem Sinne hat dieser Artikel – durch die Migration des am Fraunhofer FOKUS entwickelten GovData-Piloten in eine Cloud-Umgebung – den Grundstein für das Anbieten eines ”Urban-Platform-as-a-Service” Modells gelegt.

Literatur

Amazon. (2017, 08 24). *AWS | Amazon Elastic Compute Cloud*. Retrieved 08 24, 2017, from https://aws.amazon.com/de/ec2/

BSI. (2016). *PAS 182:2014 Smart city concept model – Guide to establishing a model for data interoperability.* Retrieved 08 24, 2017, from http://shop.bsigroup.com/upload/268968/PAS%20182_bookmarked.pdf

Definition, O. (2017, 08 24). *Open Definition. The Open Definition.* (Open Definition) Retrieved 08 24, 2017, from http://opendefinition.org

ENISA. (2017, 08 24). *Enisa European Network and Information Security Agency*. (ENISA) Retrieved 08 24, 2017, from http://www.enisa.europa.eu

JMeter. (2017, 08 24). *Apache JMeter - Apache JMeter™*. (Apache) Retrieved 08 24, 2017, from http://jmeter.apache.org/

Junar. (2017, 08 24). *Junar | The Open Data Platform*. Retrieved 08 24, 2017, from http://www.junar.com/index9ed2.html?lang=en

Liferay. (2017, 08 24). *Liferay*. (Liferay) Retrieved 08 24, 2017, from http://liferay.org

OKFN. (2017, 08 24). *CKAN. The Open Source Data Portal Software.* (OKFN) Retrieved 08 24, 2017, from http://ckan.org

PostgreSQL. (2017, 08 24). *PostgreSQL*. (PostgreSQL) Retrieved 08 24, 2017, from http://postgresql.org

RackSpace. (2017, 08 24). *Rackspace: Managed Dedicated & Cloud Computing Services*. Retrieved 08 24, 2017, from https://www.rackspace.com/de

Ramakrishnan, R., Shrawan, V., & Singh, P. (2017). Setting Realistic Think Times in Performance Testing: A Practitioner's Approach. *ISEC '17 Proceedings of the 10th Innovations in Software Engineering Conference .* Jaipur.

Scholz, R., Nikolay, T., Lämmel, P., & Schieferdecker, I. (2017). A CKAN Plugin for Data Harvesting to the Hadoop Distributed File System. *{CLOSER} 2017 - Proceedings of the 7th International Conference on Cloud Computing and Services Science.* Porto.

Tcholtchev, N., Farid, L., Marienfeld, F., Schieferdecker, I., Dittwald, B., & Lapi, E. (2012). On the interplay of open data, cloud services and network providers towards elec- tric mobility in smart cities. *In LCN Workshops*, (pp. 860-867).