# Bachelorarbeit von Viet Hoa Nguyen bei der Immowelt AG

Nürnberg, 21.03.2019

## Titel der Arbeit

Entwicklung eines verteilten cloud-basierten Systems zum Crawling von öffentlichen immobilien-relevanten Daten für ein großes deutsches Immobilienportal

Title of the thesis

Development of a distributed cloud-based system for crawling public real estate relevant data for a large German real estate portal

Motivation

Bauvorhaben, Bebauungspläne, Umwidmungen von Flurstücken etc. können drastische Änderungen im Wert der betroffenen Immobilien und Flurstücken bedeuten. Die Veränderungen können im Voraus in Erfahrung gebracht werden, z.B. durch öffentliche Protokolle der Bauausschüsse oder Amtsblätter. Diese werden aber nicht von allen Eigentümern gelesen. Die entsprechenden Webseiten aller Städte, Gemeinden, anderer Behörden und Ämter sollen daher regelmäßig gecrawlt und neue/aktualisierte Immobilienwert-relevante Dokumente gespeichert werden. Wichtig ist der effiziente Einsatz von Crawlern, d.h. Webseiten möglichst nur dann zu crawlen, wenn auch neue/aktualisierte Dokumente zu erwarten sind.

Motivation

Construction project, real estate development project and reallocation of cadastral unit can have a large impact on the value of the real estate objects. These changes can be early informed with inspecting the official announcement, remark or protocols. For this purpose the websites of all cities and administrative districts should be frequently crawled in order to retrieve up-to-date announcements relevant to real-estate objects. The application of an efficient web crawler, which periodically works on schedule when there is a new announcement released, is inevitable to accomplish this project.

Aufgaben

* Konzeption und Implementierung einer verteilten Architektur auf Basis von Cloud Computing Technologie am Beispiel des Amazon Web Service mit intelligentem Scheduling in Abhängigkeit von Erscheinungsdatum von bisherigen Amtsblättern
* Automatisiertes Deployment und Skalierung in der Cloud.
* Vorschlag eines Modells zur Steuerung und Monitoring der Web-Crawler-Worker und -Tasks
* Test und Auswertung der Funktionalität

Tasks

* Conception and implementation of a distributed web crawler that utilizes cloud computing technologies with intelligent functionalities such as scheduled crawling plan based on release date of previous official announcement.
* Automatic deployment and scaling of the crawler in the cloud
* Proposal of a model for monitoring and controlling of the crawler and crawling tasks
* Testing and evaluation of the crawler functionalities

|  |  |
| --- | --- |
| Technische Hochschule Nürnberg Georg-Simon-Ohm  Fakultät Informatik | |
| Prof. Dr. Zapf Michael Prof. Dr. Stappert Friedhelm | |
| Immowelt AG  Products Management – Big Data | |
| Axel Schwanke  Maxim Fridental | |