Bachelor thesis of Viet Hoa Nguyen at Immowelt AG

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**

Abstract

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.

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 for monitoring and controlling of the crawler and crawling tasks

Testing and evaluation of the crawler functionalities as well as the adopted AWS Services

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

Table of contents

List of Figures

List of Tables

List of Diagrams

List of Tables

List of Abbreviations

1. Introduction
   1. Immowelt AG
   2. About the project “Baukarte” (Interactive building profile)
   3. Motivation
   4. Problem statement
   5. Aim of the work
   6. Methodological approach
   7. Structure of the thesis
2. Theoretical background
   1. Web Crawler
      1. Anatomy of a web crawler
      2. Types of web crawler
      3. Incremental web crawler
      4. Challenges
   2. Cloud Computing
   3. Amazon Web Services
3. Design and Conception
   1. Requirements analysis
      1. Technical requirements
      2. Non-technical requirements
      3. The key criteria for a cloud-based solution
      4. On-demand cloud computing platforms in comparison
      5. Evaluation of candidates based on the key criteria
   2. Conception
      1. Conceptual crawler architecture
      2. Conceptual crawling process
   3. AWS and AWS Services
      1. Description of adopted services
      2. Crawler architecture on AWS
4. Implementation and realization
   1. Development environment and Frameworks
   2. Implementation of the web crawler
   3. Implementation for monitoring and scheduling
5. Conclusion
   1. Testing and evaluation of the web crawler
   2. Evaluation of the adopted AWS Services

List of Appendices

List of Cited Literature