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

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

**Structure of the thesis**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Issues** | **Definition of the objectives** | **Methods** | **Result** |
| **Abstract** |  | | | | |
| **Chapter 1: Introduction** | Project “Baukarte” | How can this thesis be positioned in the whole project? |  |  |  |
| Motivation | How can official announcement be effectively collected with cloud-based web crawling? |  |  |  |
|  |  |  |  |  |
| **Chapter 2:**  **Theoretical Background** | Anatomy of a web crawler | In which principle does a web crawler work? |  | Research, Literature analysis | Analysis of a general *web crawling algorithm* |
| Incremental web crawler | Which type of web crawler is suitable for the use case at *Immowelt AG* ? |  | Research, Literature analysis | Early *recognition* of *technical and non-technical requirements*  Early *identification of difficulties* that a web crawler often deals with |
| Cloud computing | What is *cloud computing* and what is its *benefits*? | Overview about cloud computing |  | Introduction and description of *terms* in *cloud computing* |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Issues** | **Definition of the objectives** | **Methods** | **Result** |
|  | 2.4 Amazon Web Service | How does *Amazon Web Services* work and its advantages in compare with other cloud-computing platform? | Possibilities to use AWS Services on current use case | Product research | AWS Services to be adopted |
| **Chapter 3:**  **Design and Conception** | Requirement analysis | How should the web crawler be implemented and managed? | Identification of requirements for a web crawler in terms of functions and customized specification from *Immowelt AG* | Requirement analysis and engineering | A *study* based on evaluation of existing commercial web crawler  Recommendation and best practice for implementing web crawler |
| Conception | What is the main component? | Plan for implementation | Software engineering | *Initial architecture* for cloud-based web crawler that satisfy technical as well as non-technical requirements |
| AWS and AWS Services | Which AWS Services can be utilized in this current scenario? | Study of existing suitable services from AWS in term of price and availability | Cost-benefit analysis  ABC Analysis of different AWS Services | *Initial plan* of AWS Services being used |
| **Chapter 4**  **Implementation and realization** |  | How can a cloud-based web crawler be implemented, managed and controlled?  How far can the initial architecture and service plan proposal be implemented? | Development of a Web Crawler using Amazon Web Services by reference to initial architecture | Proof of Concept  Prototyping | *Web Crawler Prototype* including  - Crawling pipeline  - Accomplished implementation of a web crawler in python  - Proposal for management and deployment process |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Issues** | **Definition of the objectives** | **Methods** | **Result** |
| **Chapter 5:**  **Conclusion** | 5.1 Evaluation of application in this thesis | Which results have been created in this thesis and how performant does this work?  Which result could not be achieved? | Achievements of the protype | Evaluating  Validation | *Comparison table* between early requirement and actual implementation  *Advantages* and *drawbacks* of the prototype |
| 5.2 Evaluation of AWS | Which are the advantages and disadvantages of cloud-based solution? |  | Lesson-learned |  |
| 5.3 Future work | Which improvements to this thesis can be done |  | Requirements engineering | Recommendations for using AWS Services and web crawling in general |