EE/CprE/SE 491 - sddec22-19
Text extractions from documents into Elasticsearch
Week 2 Report

2/7/22 - 2/13/22
Client: Kayla Gilleland, Buildertrend
Faculty Advisor: Goce Trajcevski

Team Members:

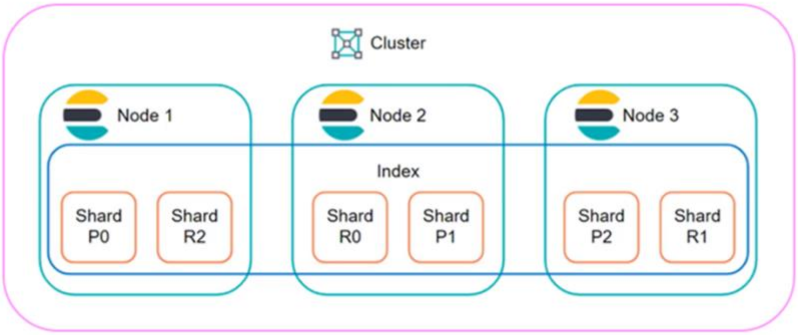
Bruce Bitwayiki
Jared Hayashi
Rushal Sohal
Tiffany Mayberry

Weekly Summary

The team met with the client this week. We were able to ask questions regarding the project scope as well as additional technical questions(files support, technology stack, goals and deliverables, etc). Client briefly discussed the expectations for this semester and how we will be communicating with them (MS Teams). The team took part in a group TeamThink activity to discuss individual work styles and better understand how their workstyle, strengths and weaknesses can contribute to the group. The team also established an official team contract to which all members must abide by and had it signed and submitted. The contract outlines procedures, expectations, member roles, and consequences for not adhering to the contract.

Past Week Accomplishments

- Resolved questions on project scope - All team members
  - Project is more of a proof of concept, and we will be starting from scratch.
  - Word documents, pdfs, powerpoint, and excel documents are the important file types for now
  - UI platform is flexible, and will be used to show the functionality of our application
- Elasticsearch Research - Tiffany
  - Looked into Elasticsearch functionality and features. Highlighted below are
    - Use SQL queries or HTTP request to insert data (individual or bulk) and query data
    - Handles indexing and allows fast queries
    - Cluster design eases scalability and reliability concerns



Node: A single instance of Elasticsearch

Cluster: A collection of nodes, working together

Index: A container for documents within the cluster, consists of shards

Shard: Unit of work, primary and replica shards

Document: Data structure to store information (JSON), belongs to an index

- Apache Tika Research
- Lucene Research
- Worked on the team initiation assignment
- Start Documentation
- SWOT analysis:

SWOT Analysis for the New Search Functionality	
Strengths: • Allow access to more information that has been uploaded • Better experience for users	**Weakness:** • Increase data storage requirements for documents • Cost of elasticsearch
Opportunity: • Application in new areas • Improve searching	**Threats**: • Security concerns (data leak etc.) • No support for obscure file types

Individuals Contributions

Name	**Individual Contributions**	**Hours This Week**	**Hours Cumulative**
Bruce Bitwayiki	Tech Stack research(Elasticsearch tutorial, TIKA), Documentation(format, reqs., template)	3	4
Jared Hayashi	Apache Tika Research	1	2

Rushal Sohal	Researching (lucene, tika) and documentation	1	2
Tiffany Mayberry	Elasticsearch Research	4	5

Plans for the Upcoming Week

- Get in touch with Kayla to get a Teams room set up for easier communication
- Ask instructor about what forms we need to fill out (NDA, professionalism, etc.)
- Get task management tools setup for future use (Gitlab milestones)
- Start developing framework for final design document in Google Drive
- Begin documentation (Problem statement, Reqs., Constraints, Users and uses, Project Plan, etc)

Summary of Weekly Advisor Meeting

- Reviewed information from kickoff meeting on Monday
- Went over forms that need to be filled out prior to working with an industry partner (NDA, professionalism, etc.)
- Looked over completed form for sdmay22 in order to get a feel for the type of work that will need to be done this semester