Text Extractions from Documents into Elasticsearch

Team: sddec22-19

Client: Buildertrend **Advisor:** Goce Trajcevski

Team Members: Bruce Bitwayiki, Jared Hayashi, Rushal Sohal, Tiffany Mayberry

http://sddec22-19.sd.ece.iastate.edu/

Overview

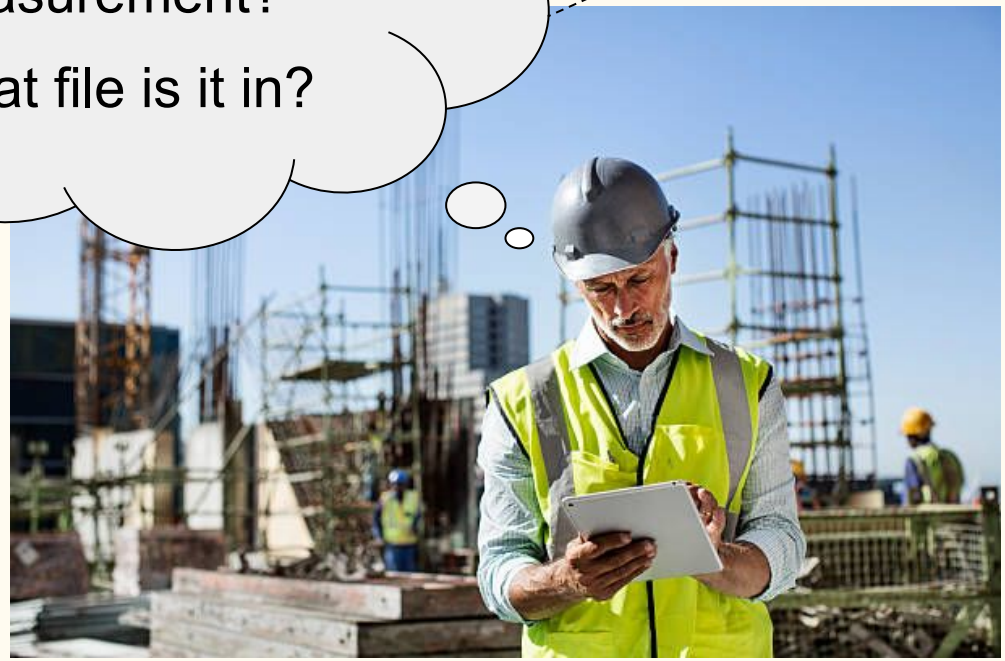
The Client

- Industry-Leader in Construction Management Software
- An all-in-one platform: set a budget, communicate with clients, schedule work and keep tabs on job site progress
    - Presales, Project Management, Financial Tools, Customer Management
- 27,000+ construction companies use Buildertrend

The Problem

What is the measurement?

What file is it in?



BUILDERTREND

Functional Requirements

Upload

Extract File Content & Metadata

JPG PNG PDF .TXT

Search for Files Containing:
"5213 Pines Dr."

Filters:
Only search .docx, .jpg, .xlsx files

Non-Functional Requirements

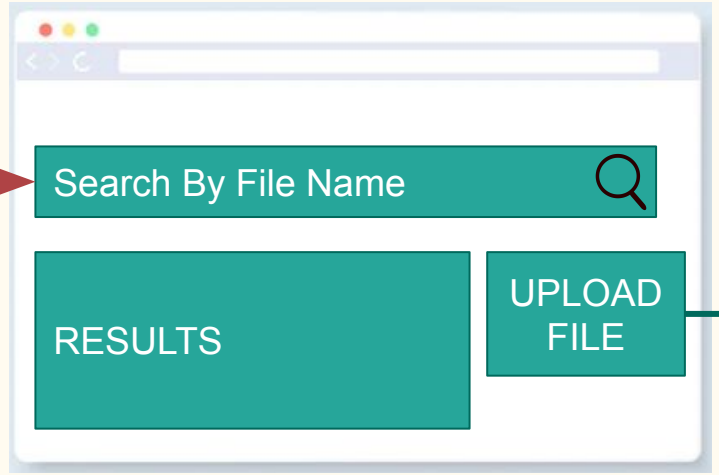
Easy to use UI

Return result within 10 seconds

Easily modifiable and scalable for future functionality

Design

Buildertrends Current Solution

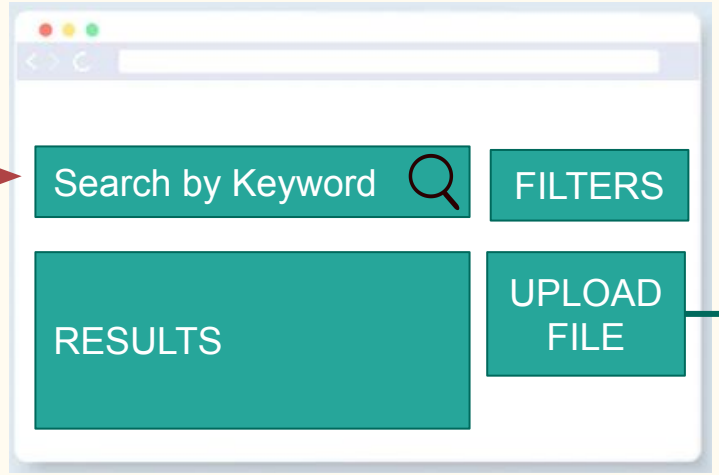


Fetch & Return Results

Upload File

DATABASE

Basic Overview of Our Project



Fetch & Return Results

EXTRACT TEXT & INSERT INTO DATABASE

elasticsearch

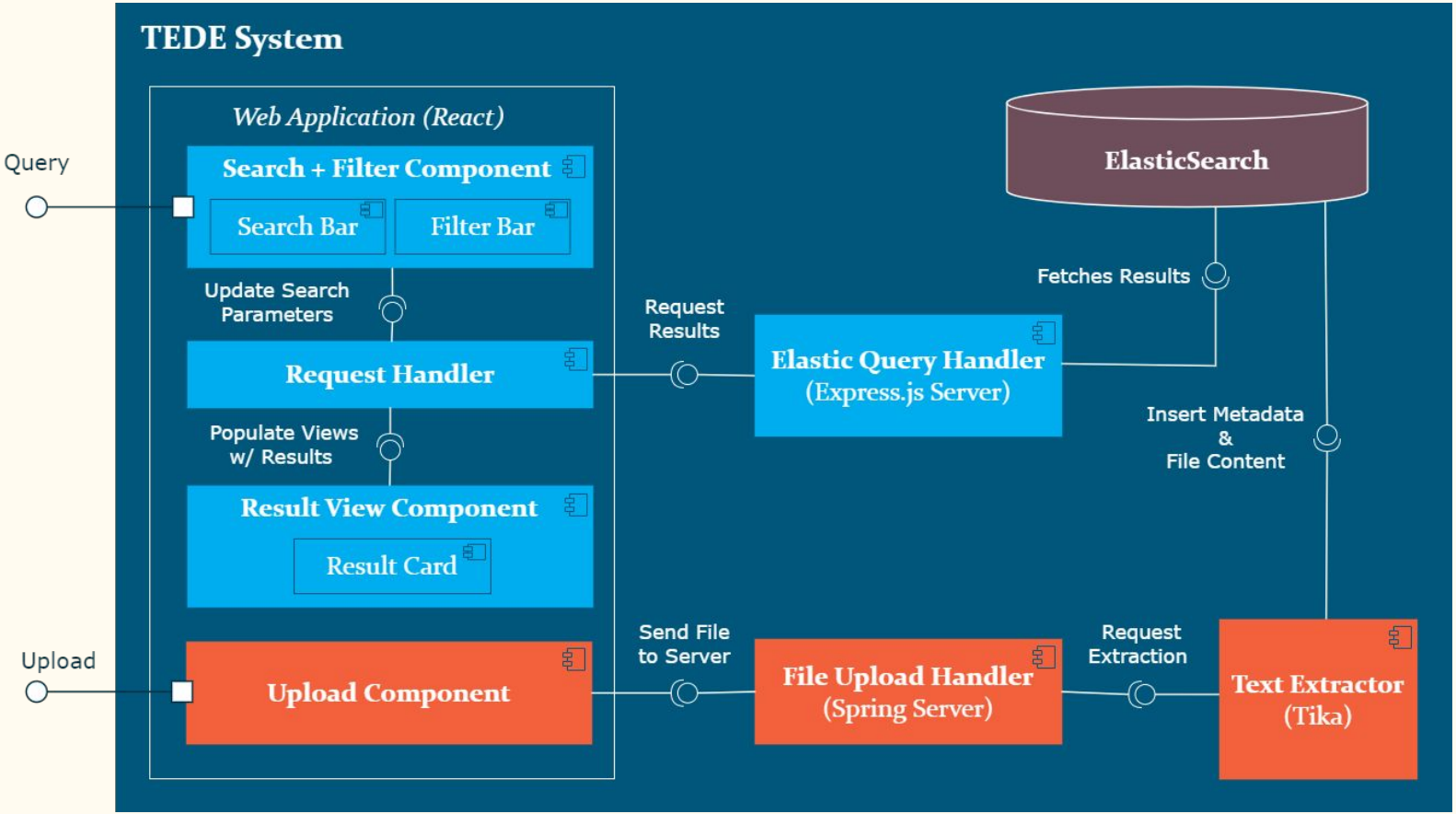
(our database)

Technology Stack

Web Application	Database	Text Extractor
		
		

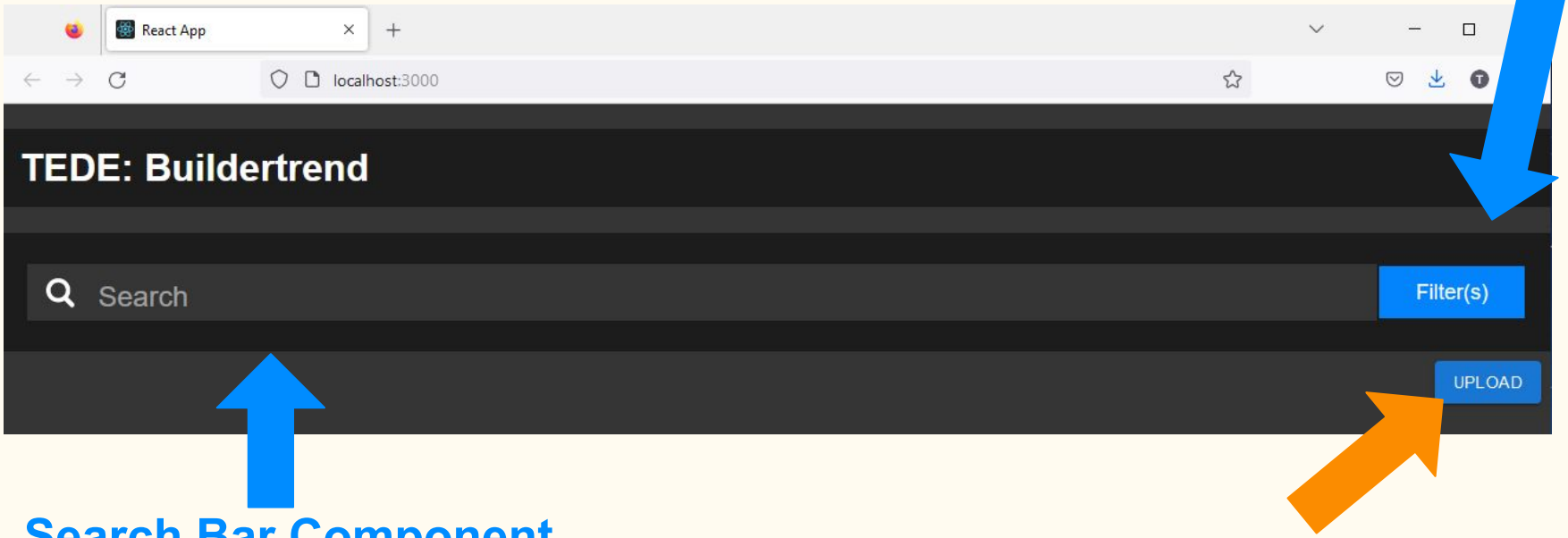
Implementation

Component Diagram



Frontend Start Page

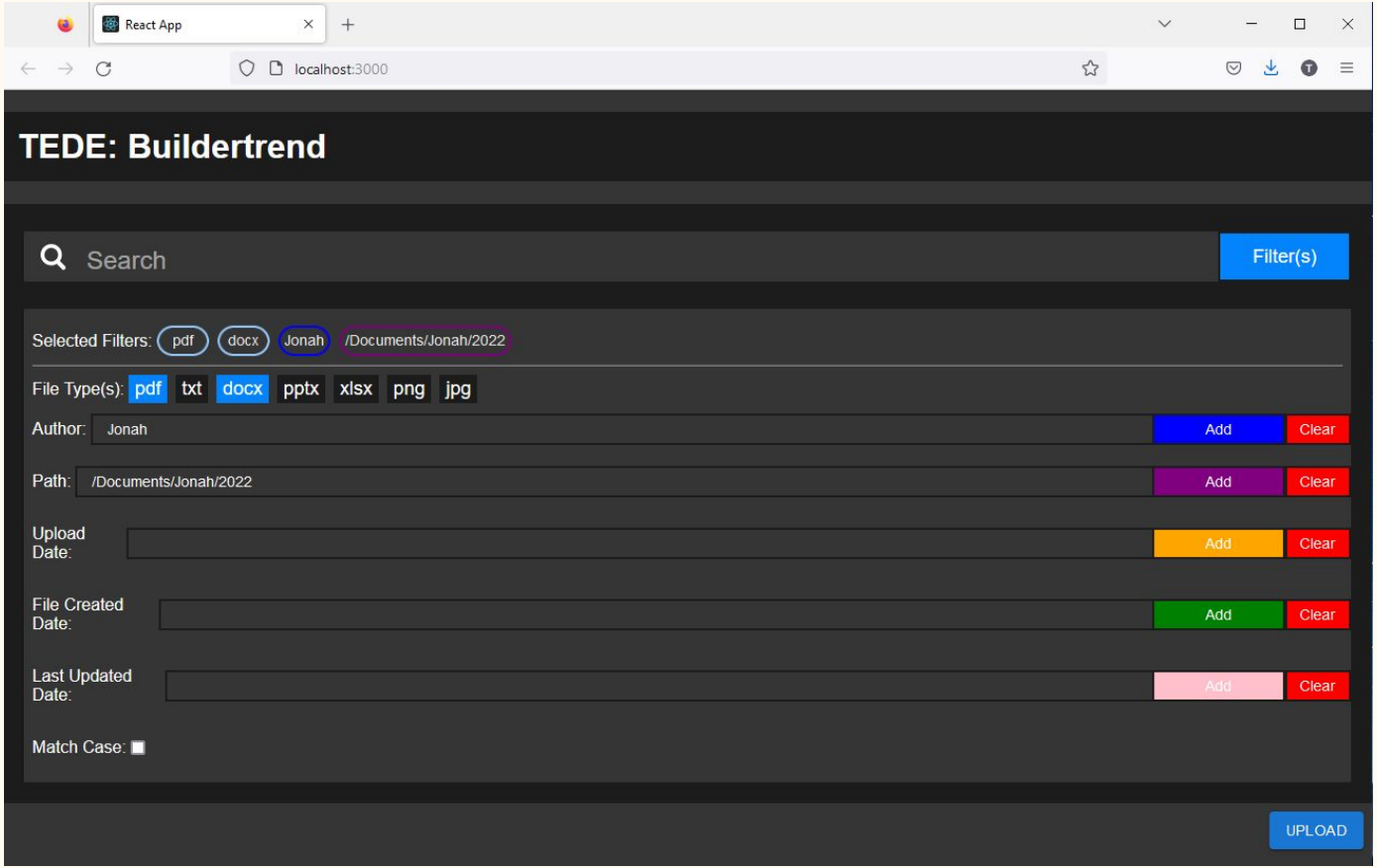
Filter Bar Component
Shows available filters



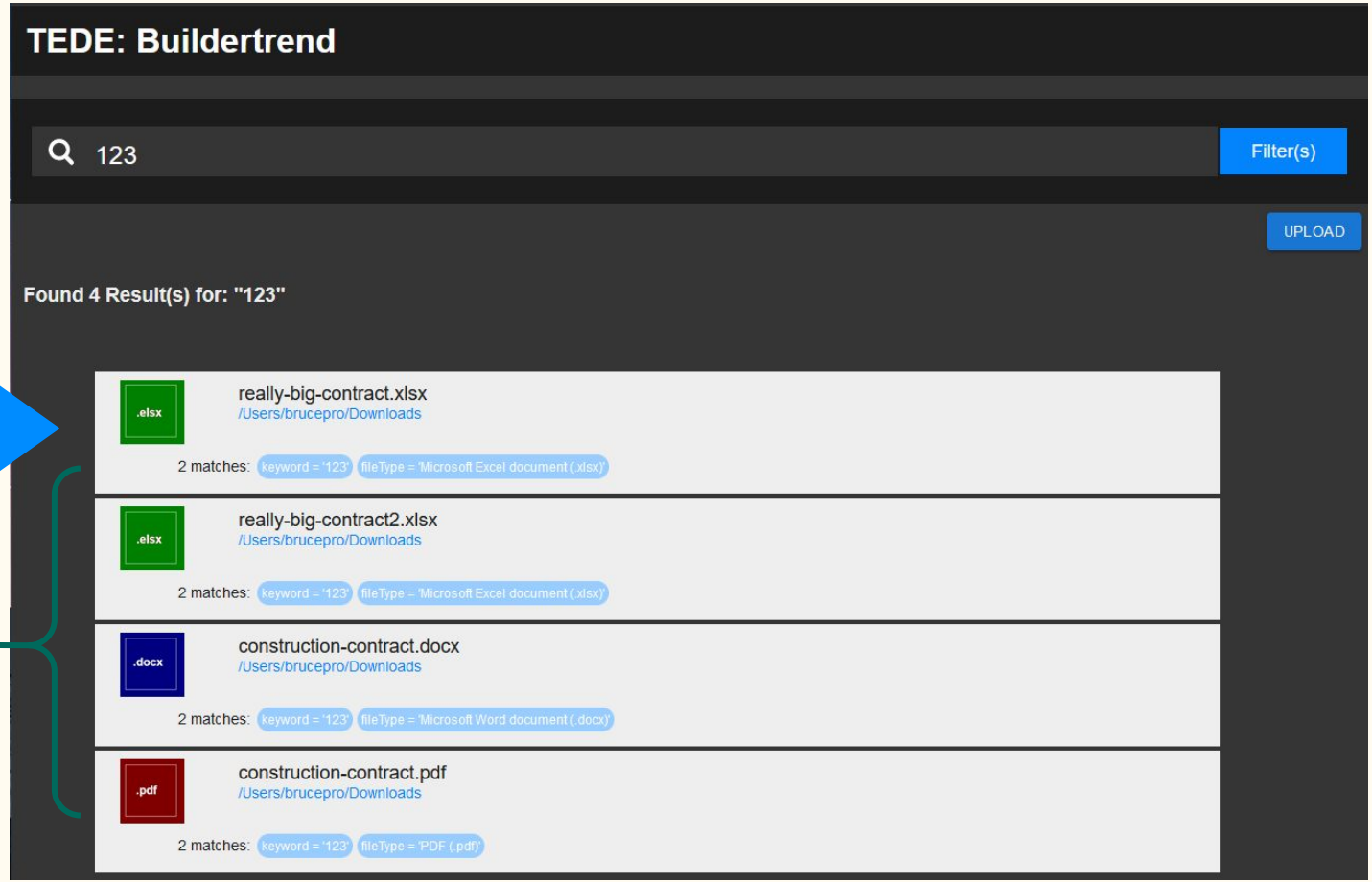
Search Bar Component
Sends inputs to Request Handler Component (on enter)

Upload Component
Gives user the ability to upload supported file types

Frontend Expanded Filter Bar



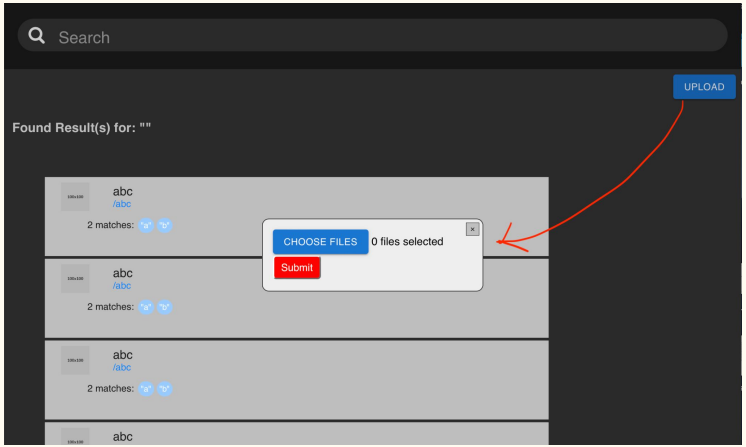
Frontend Results View

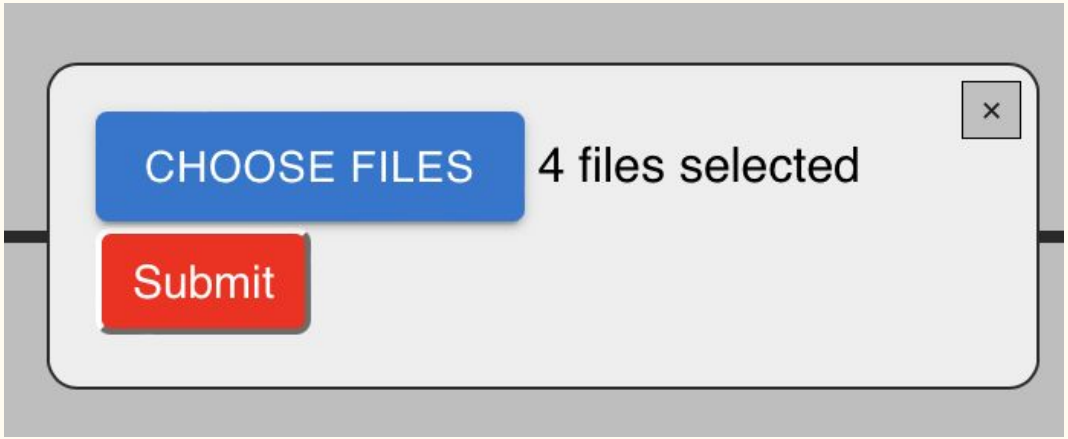


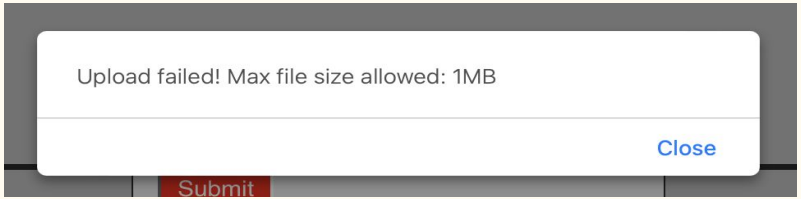
Result Card

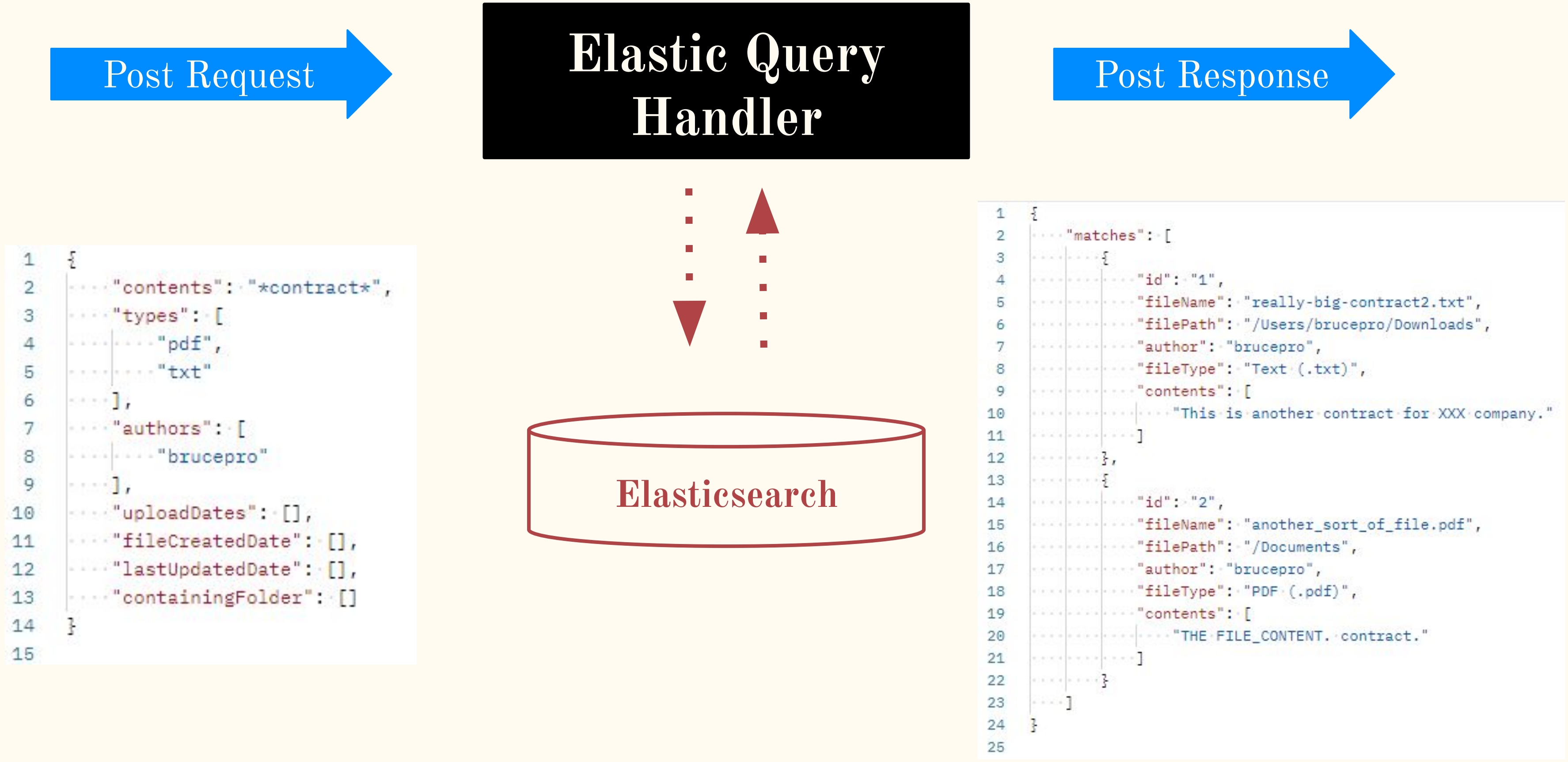
Result View

Upload Component








Post Request
Elastic Query Handler
Post Response
Elasticsearch
{
"contents": "*contract*",
"types": [
"pdf",
"txt"
],
"authors": [
"brucepro"
],
"uploadDates": [],
"fileCreatedDate": [],
"lastUpdatedDate": [],
"containingFolder": []
}
{
"matches": [
{
"id": "1",
"fileName": "really-big-contract2.txt",
"filePath": "/Users/brucepro/Downloads",
"author": "brucepro",
"fileType": "Text (.txt)",
"contents": [
"This is another contract for XXX company."
]
},
{
"id": "2",
"fileName": "another_sort_of_file.pdf",
"filePath": "/Documents",
"author": "brucepro",
"fileType": "PDF (.pdf)",
"contents": [
"THE FILE_CONTENT. contract."
]
}
]
}

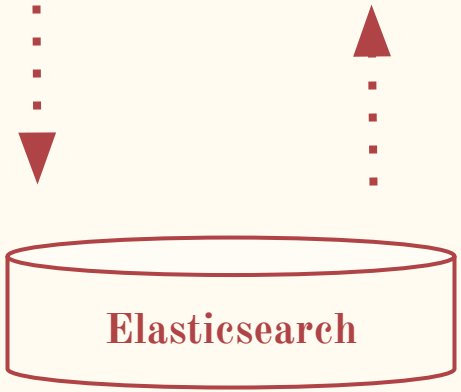
Post Request

Elastic Query Handler

Post Response

Elasticsearch Query

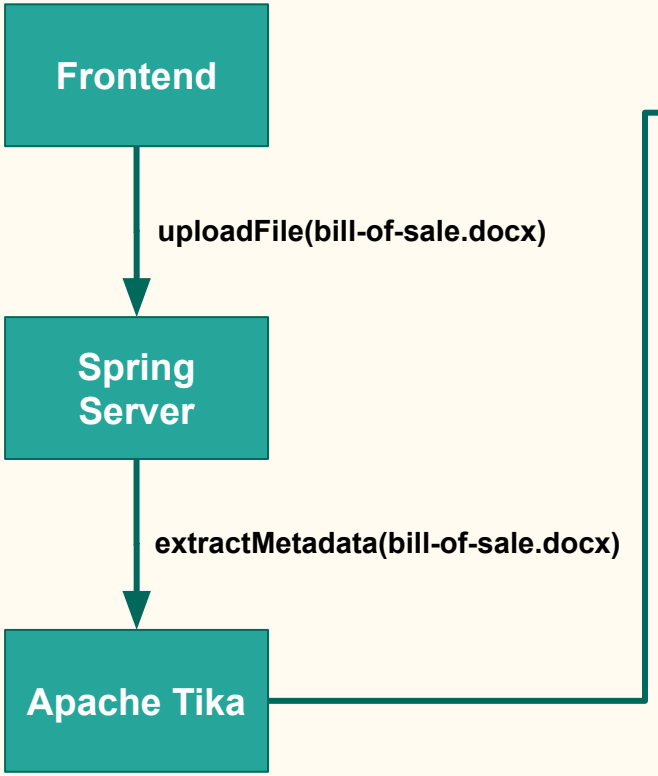
```
{
    index: 'documents',
    query: {
        bool: {
            "must": [
                { "query_string": {
                    "default_field": "contents",
                    "query": "*contract*"
                }}

            ],
            "filter": [
                { "bool": {"should": [
                    {"match": { "type": "PDF (.pdf)" }},
                    {"match": { "type": "Text (.txt)" }}
                ]}},
                { "bool": {"should": [
                    { "match_phrase": { "owner": "brucepro" } }
                ]}}
            ]
        }
    }
}
```



Elasticsearch Response

```
{
    query: params.body.query,
    hits: { hits:[
        {
            _id: "1",
            "_source" : {
                "type" : "Text (.txt)",
                "title" : "really-big-contract2.txt",
                "path" : "/Users/brucepro/Downloads",
                "owner" : "brucepro",
                "contents" : "This is another contract for XXX company."
            }
        },
        {
            _id: "2",
            "_source" : {
                "type" : "PDF (.pdf)",
                "title" : "another_sort_of_file.pdf",
                "path" : "/Documents",
                "owner" : "brucepro",
                "contents" : "THE FILE_CONTENT. contract."
            }
        }
    ]
    }
}
```

Backend Example



JSON sent to Elasticsearch


```
{
    "owner": "Microsoft Office User",
    "path": "/home/vm-user/sddec22-19/ExampleDocs/bill-of-sale.docx",
    "creator": "Microsoft Office User",
    "contents": "Bill of Sale THIS BILL OF SALE is executed this day of _____, by _____, (hereinafter
Seller) residing at _____, _____ _____ for the benefit of
_____ (hereinafter Buyer), residing at _____, _____, _____ _____
_____. Seller hereby transfers to Buyer, all rights of Seller in the following property: _____, located in _
_____ County, _____. For and in consideration of _____, which has been acknowledged
to have been received by Seller. The form of payment used will be _____ and sales tax will not be included as part of the
purchase price. The sale and transfer of property is made on an AS IS basis, without any express or implied warranties, with no recourse to
the Seller, provided that Seller can issue proof that it has title to the property without any liens or encumbrances. The Buyer has been giv
en the opportunity to inspect, or have inspected, any and all property as defined above. The Buyer agrees to accept all property in its exis
ting state.  In witness, the parties execute on this Bill of Sale on _____,  Signature of Buyer _____
_____ Date _____ Signature of Seller _____ Date _____
_____  Additional Notes: 1. Make sure that this Bill of Sale document is completed, and signed by both parties. Once signed, it will go
into effect on the effective date specified in the document. 1. The buyer should be provided with the original document, and a copy should b
e made and provided to the seller. 1. This document cannot be used to legally buy or sell real estate.    ",
    "file-size": "21006 bytes",
    "word-count": "284",
    "title": "bill-of-sale.docx",
    "type": "docx",
    "page-count": "2"
}
```

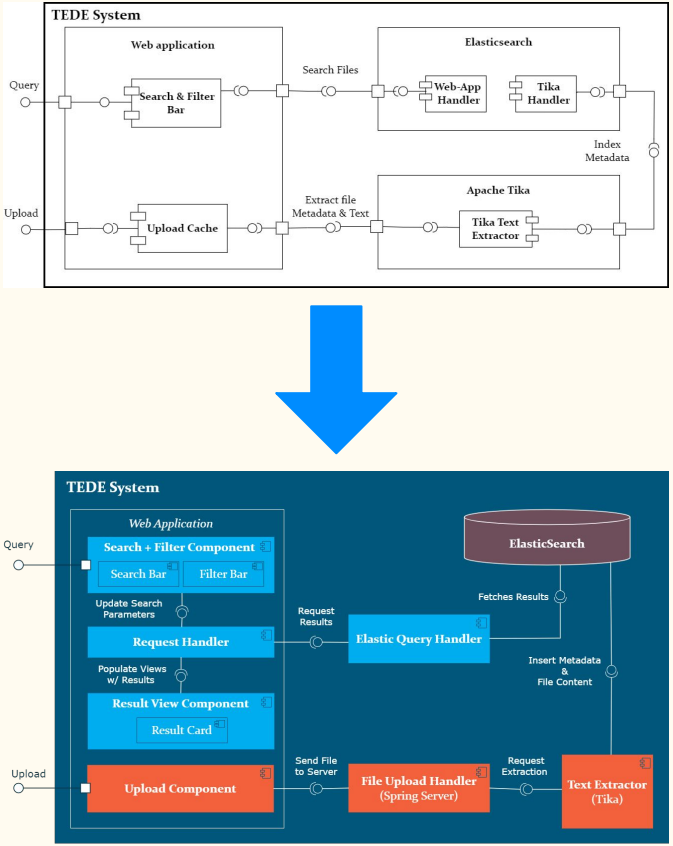


```
{
  "took" : 801,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1,
      "relation" : "eq"
    },
    "max_score" : 1.0,
    "hits" : [
      {
        "_index" : "documents",
        "_id" : "20e83493-98bc-49a1-8ecf-03dbf991d99d",
        "_score" : 1.0,
        "_ignored" : [
          "contents.keyword"
        ],
        "_source" : {
          "owner" : "Microsoft Office User",
          "path" : "/home/vm-user/sddec22-19/ExampleDocs/bill-of-sale.docx",
          "creator" : "Microsoft Office User",
          "contents" : "Bill of Sale THIS BILL OF SALE is executed this day of
```


Elasticsearch index

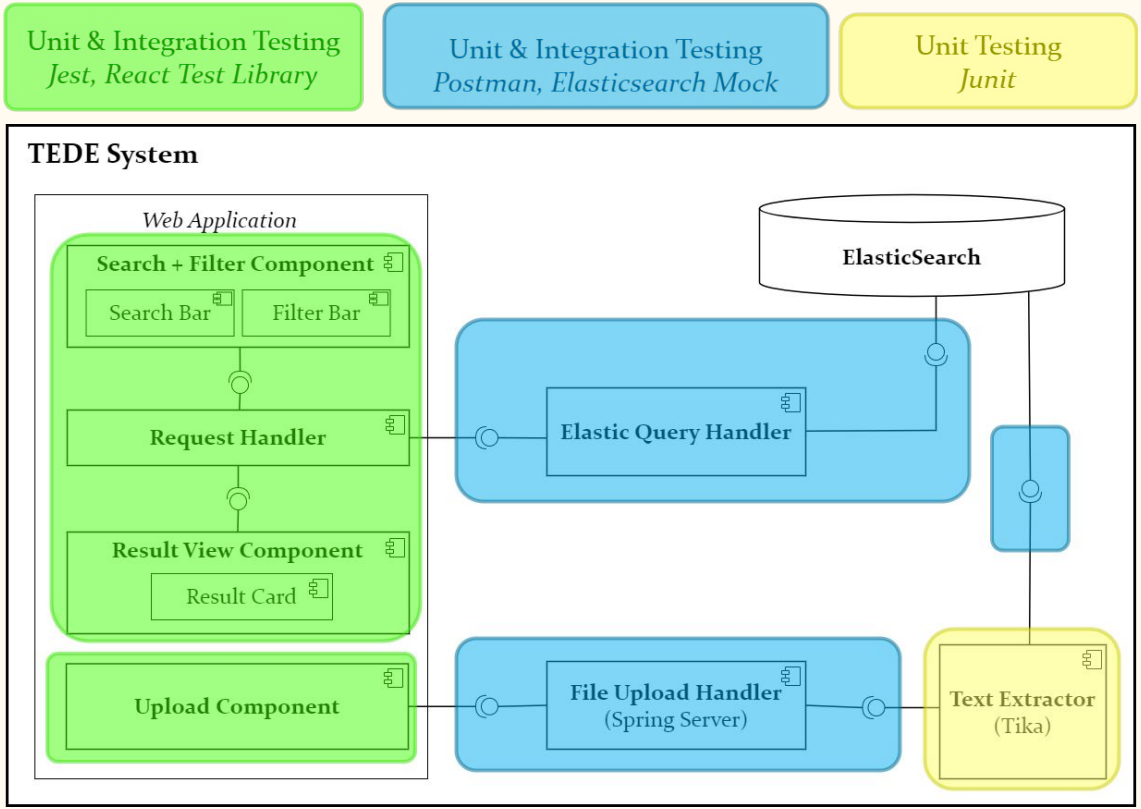
Implementation Challenges

- Elasticsearch authentication and communication
  - Certificate errors
  - Outdated documentation
- Architectural and requirement changes
  - Reallocating time to develop Tika server
  - Allocating more time for integration



Testing

Testing Plan



Acceptance Testing During BiWeekly Meetings with Client

Demo

Demo

Video

Questions ?

Extra Supporting Slides

Task Responsibility

FRONT-END

Tiffany Mayberry (SE)

Rushal Sohal (CPR E)

BACK-END

Jared Hayashi (SE)

Bruce Bitwayiki (CPR E)

Constraints and Considerations

Frontend should be a **web or desktop application**

Use **Elasticsearch** to store the metadata

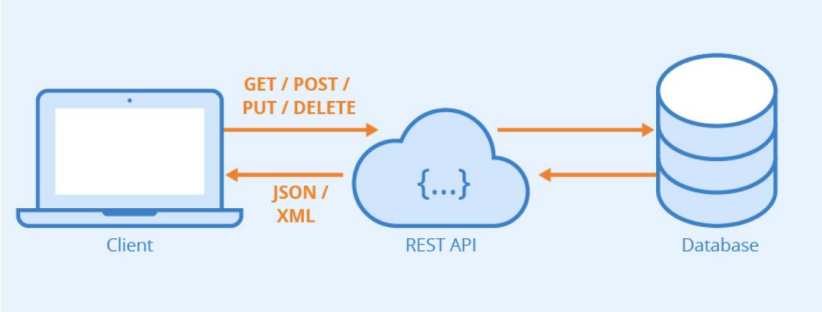
Application should **limit** the returned results

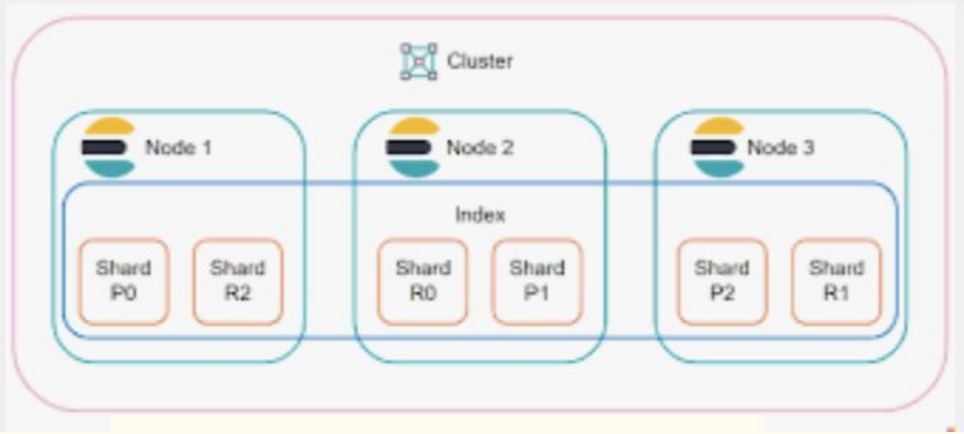
Elasticsearch

- **Free and open search and analytics engine**
- Based on the Lucene library
- Elastic Stack
- **At core - process JSON requests and return JSON data**
- **Index based searching**
- Structure based on documents
- Netflix, Ebay, Walmart, etc.

Risk Management & Mitigation

- Project Planning
- UI Integration
- REST API
- Elasticsearch Node reliability
- Text Extraction handler
- Testing





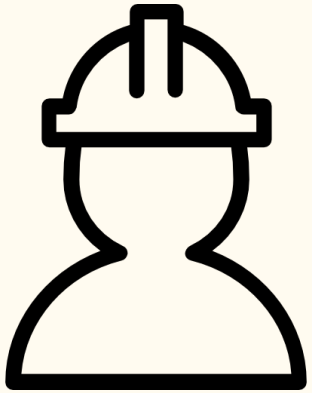
(⇧ Elasticsearch Structure)

Resource and Cost Estimate

ETG Server

Open Source Packages

Users & Needs

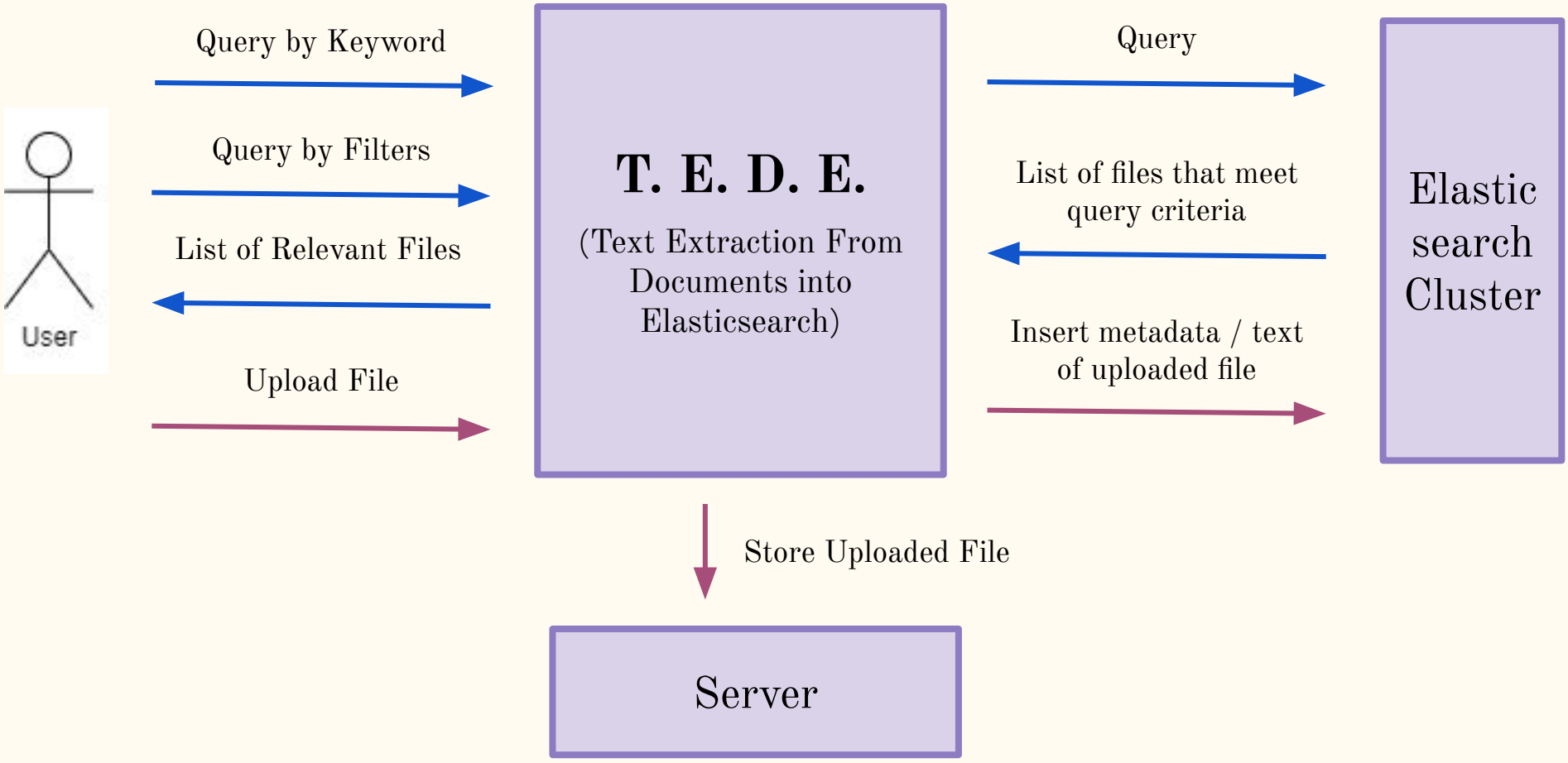


Builders / Contractors
Access to measurements, designs, orders

Homeowners
Contracts, estimates/quotes, specifications, notes

Buildertrend Staff
Support

Context Diagram



Original Testing Plan

Interface Testing
Postman

Unit Testing
Junit, Jest, React Test Library

Integration & System Testing