WARC PORTAL

CHENG CHEN | KEVIN TANG | MATE VERUNICA | ADRIANO MARINI

CMPUT 401 F16

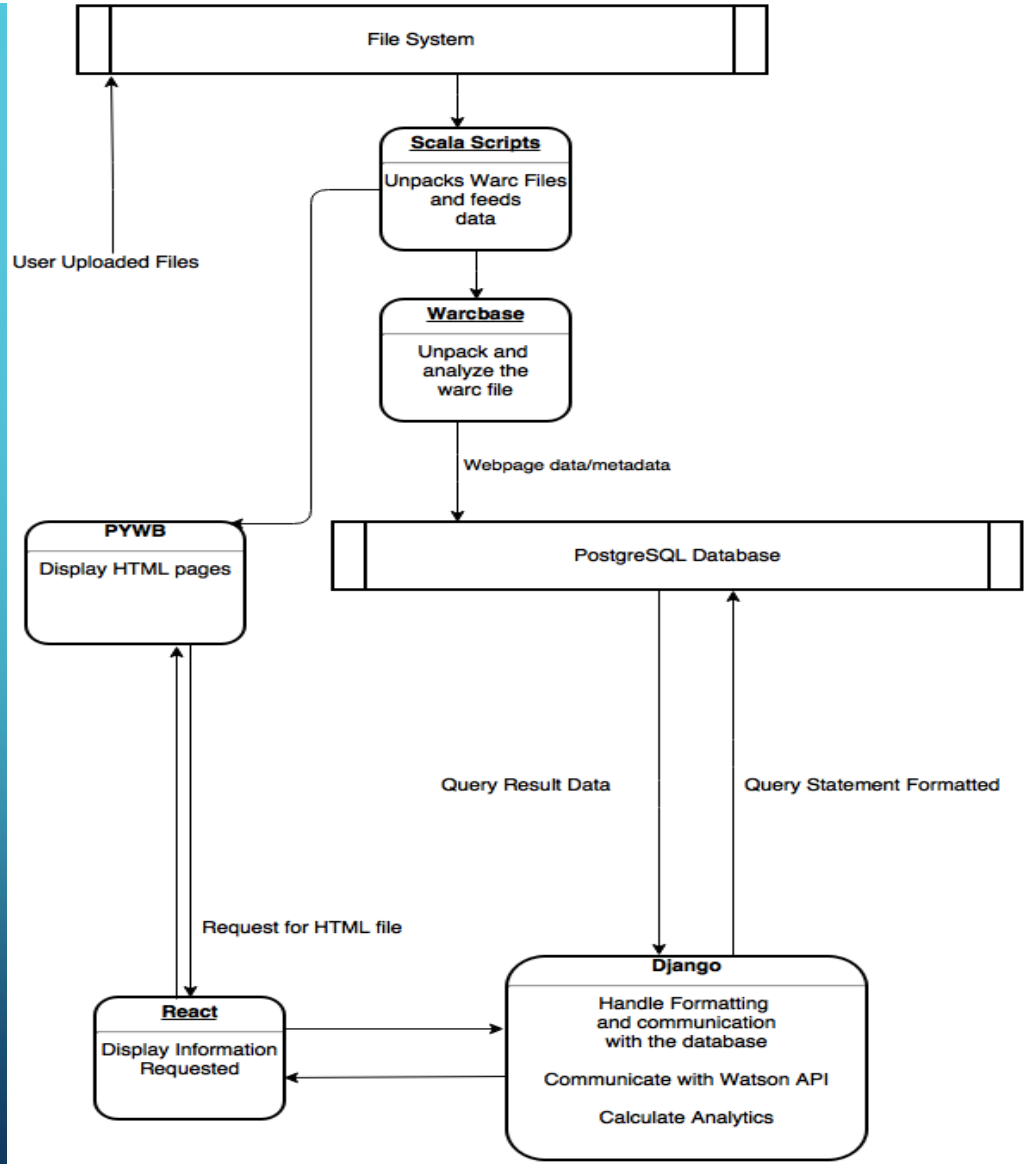
THE WARC PORTAL PROJECT AIMS TO DEAL WITH EXTRACTING, SEARCHING AND ANALYZING WEB ARCHIVE FILES. WE PROVIDE INTUITIVE AND EASY ACCESS FOR RESEARCHERS TO BROWSE AND SEARCH THROUGH THOUSANDS OF POSSIBLY DUPLICATED WEBPAGES, PROVIDE TOOLS FOR ANALYZING THEIR COLLECTIONS USING AN ARRAY OF SEARCHES AND FILTERS, AND PROVIDE HELPFUL VISUALIZATIONS OF THEIR DATA BY ANALYZING KEYWORDS.

PURPOSE AND VALUE

- Researchers have crawled web archives, called WARCs
- Disassembling and analyzing them can be a chore; We have automation!
- Provides insights into a body of data that might be collected related to specific sources
  - Use Case: Analyzing crawled articles related to the Fort McMurray Fires in 2016

ARCHITECTURE: OVERVIEW

- Three main elements:
  - Front End UI
  - Back End Server
  - Packages for Analysis and Display


File System
Scala Scripts
Unpacks Warc Files and feeds data
User Uploaded Files
Warcbase
Unpack and analyze the warc file
Webpage data/metadata
PYWB
Display HTML pages
PostgreSQL Database
Query Result Data
Query Statement Formatted
Request for HTML file
Django
Handle Formatting and communication with the database
Communicate with Watson API
Calculate Analytics
React
Display Information Requested

ARCHITECTURE: FRONT END

- Using the React Framework
  - With added Redux
- Creates the search and display framework
- Plays nicely with Django

ARCHITECTURE: BACK END

- Django server
- In combination with a PostgreSQL Database
- Implements:
  - TF-IDF Visualizations
  - Collection management
  - Talking to Watson's API for insights into documents
  - Among other standard functions

ARCHITECTURE: OTHER PACKAGES

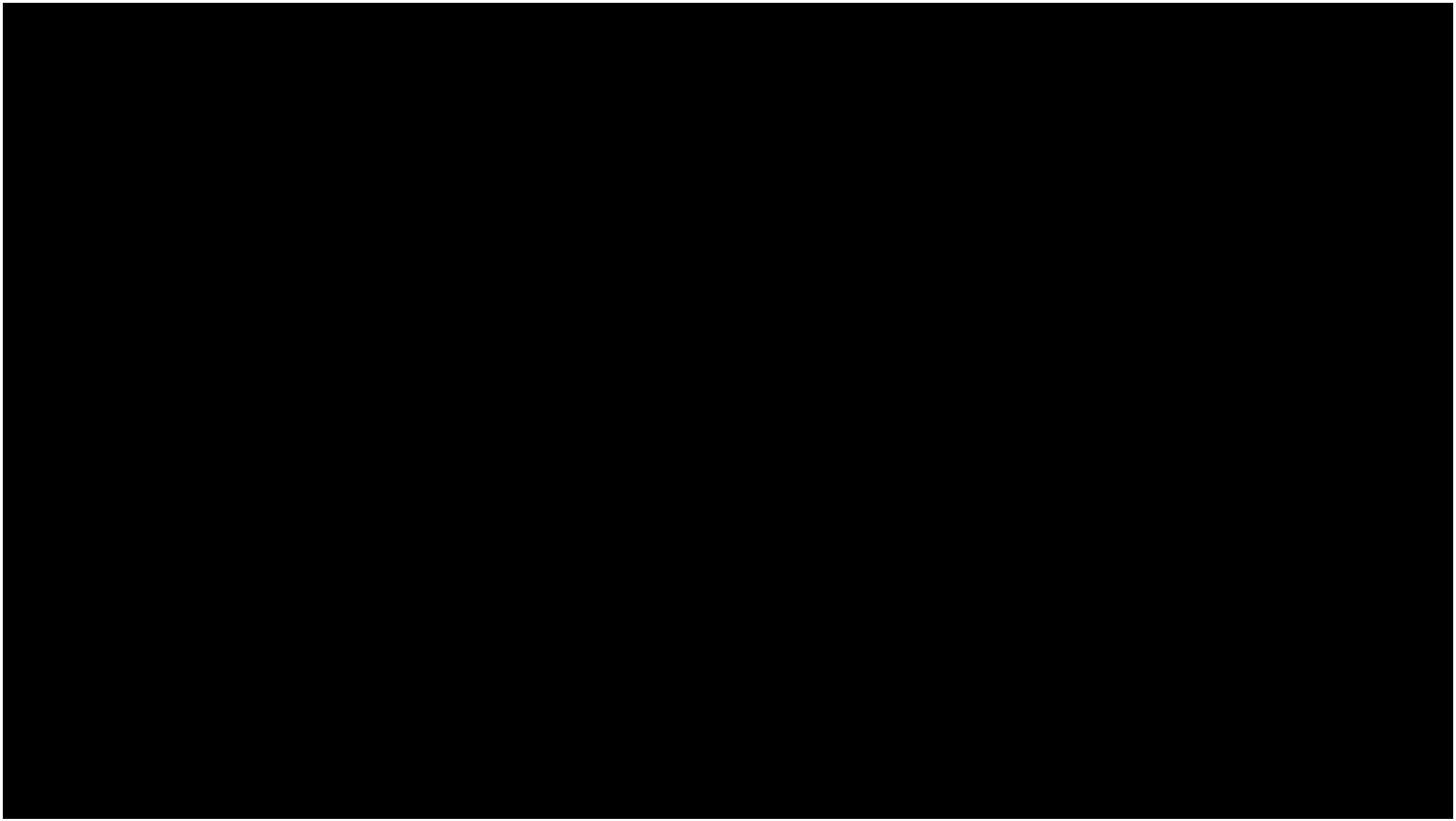
- WARCbase
  - Developed by CS @ Waterloo
  - Running on top of Apache Spark, uses Hadoop and Scala to analyze WARC files.
  - Scala scripts feed files into WARCbase for analysis
  - Good level of Automation

ARCHITECTURE: OTHER PACKAGES

- Pywb
  - Open Source package to view the archived webpages
  - Customized UI
  - Builds its own indexes on the files

ARCHITECTURE: OTHER PACKAGES

- D3
  - Creates word-cloud visualizations
- TF-IDF Ranking
  - CRON scripts
  - Automate file crawling and intake



KEY CHALLENGES

- Lack of documentation. . .
- Understanding a complex architecture
- Dirty / Bad data
- Resource restrictions
- Language / Technology Support
- MySQL's lack of Unicode Support