VIRGINIA TECH
BLACKSBURG
CS 4624

IDEAL Pages

MUSTAFA ALY & GASPER GULOTTA

CLIENT: MOHAMED MAGDY

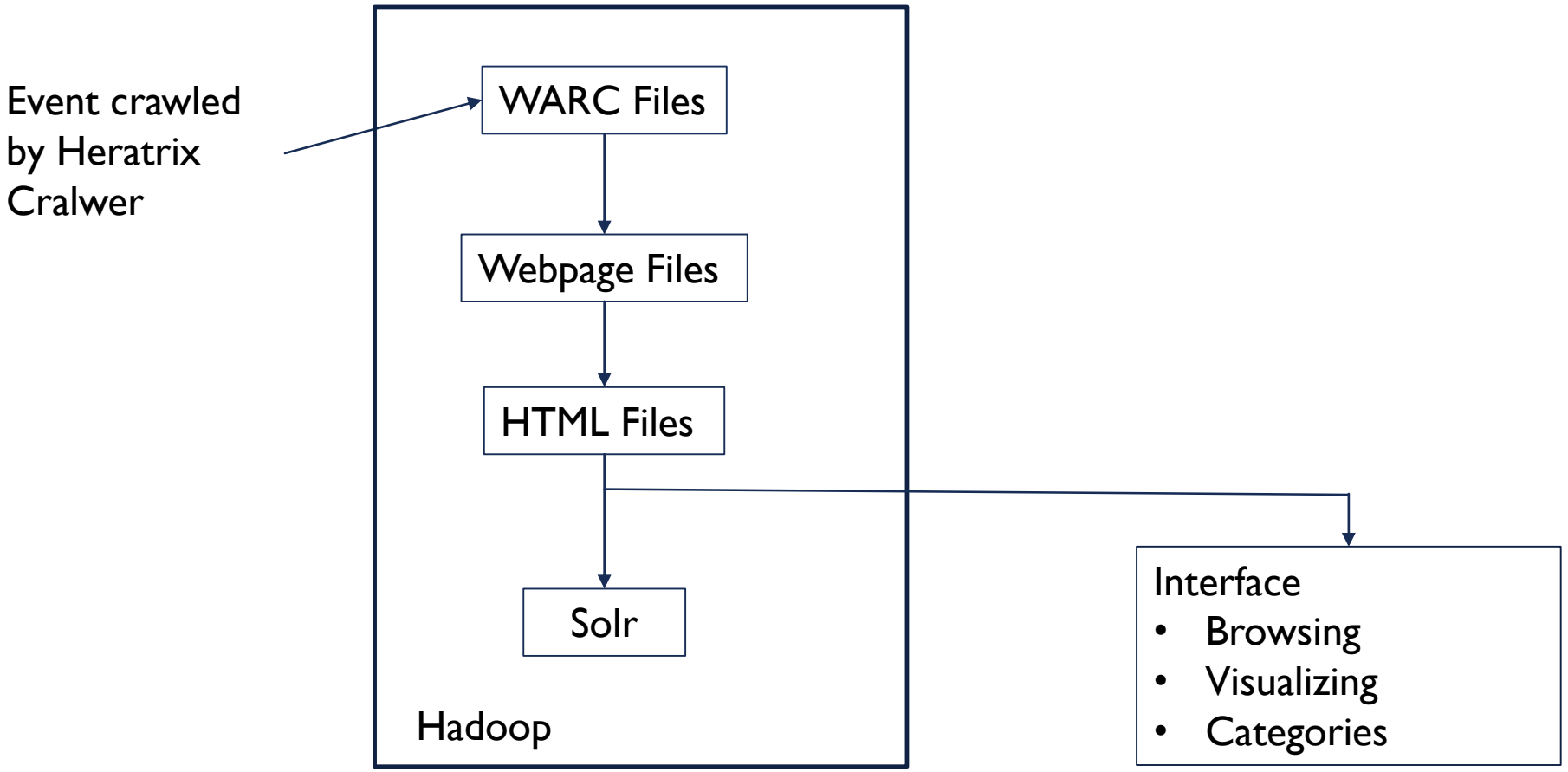
BACKGROUND

- The IDEAL Project aims to provide convenient access to webpages related to various types of disasters
- Currently this information is stored in about 10TB of Web Archives
- Need to extract this information efficiently and index it
- Provide a user interface for easy to use access

SOLUTION APPROACH

- Automate the process of:
  - Extracting the Web Archives (.warc files)
  - HTML parsing and indexing into Solr
- Use of Hadoop for distributed processing
- Webpages for displaying Solr search results and sorting disasters by category

PROJECT ARCHITECTURE



OUR ROLES

- .warc file extraction
- Filtering of HTML files
- Text extracting from HTML files
- Indexing information into Solr

WORK COMPLETED

- Set up Python environment
- Obtained a set of test .warc files
- Simplified the process of extracting a .warc file
- Identification of HTML files from the resulting extraction

WORK REMAINING

- Expand process of extracting .warc files to multiple files/directories
- Extracting text from HTML files
- Indexing information into Solr
- Work with 6604 students to integrate process with Hadoop and develop User Interface

QUESTIONS?