VIRGINIA TECH
BLACKSBURG
CS 4624

IDEAL Pages

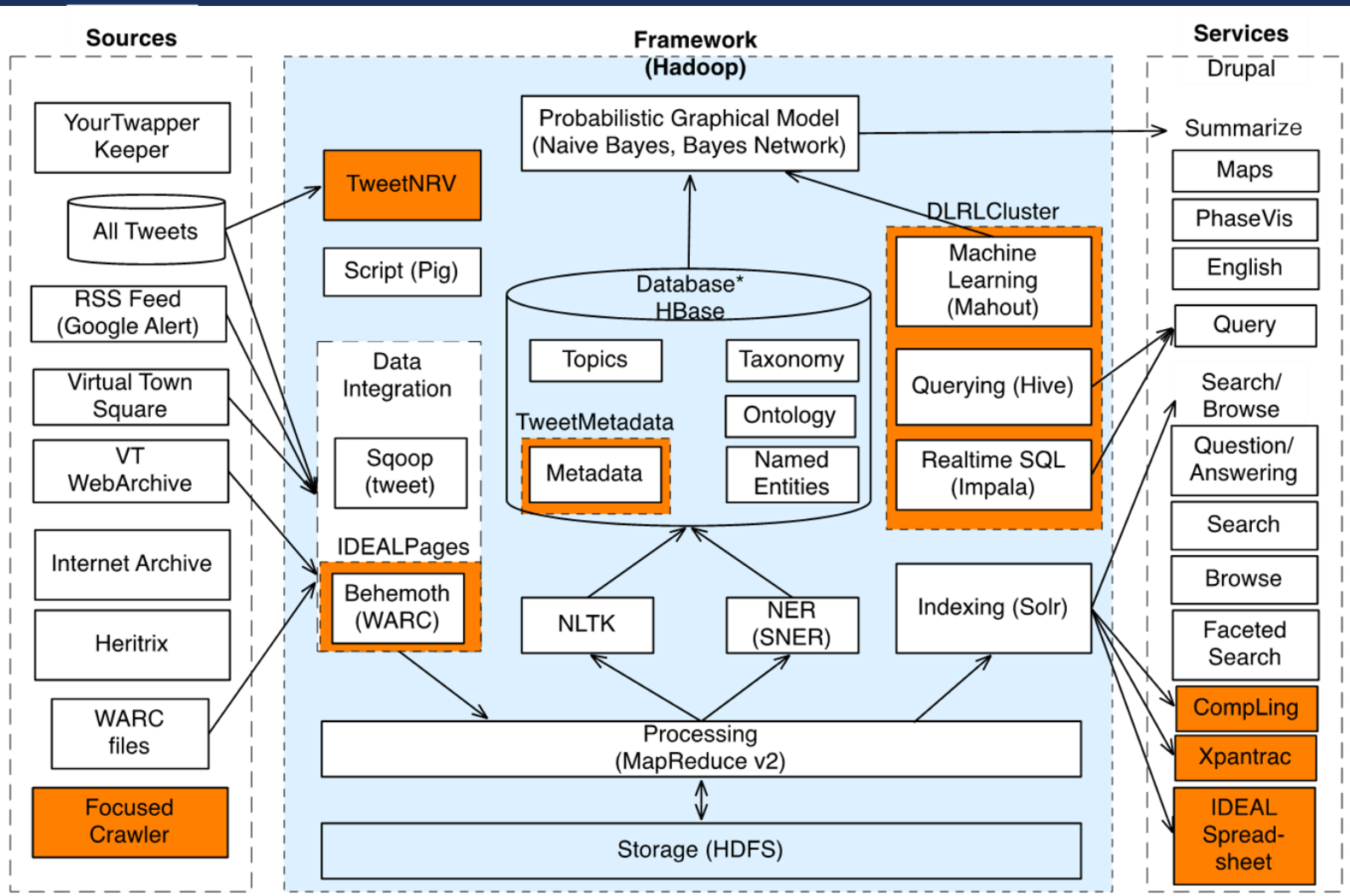
MUSTAFA ALY & GASPER GULOTTA

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

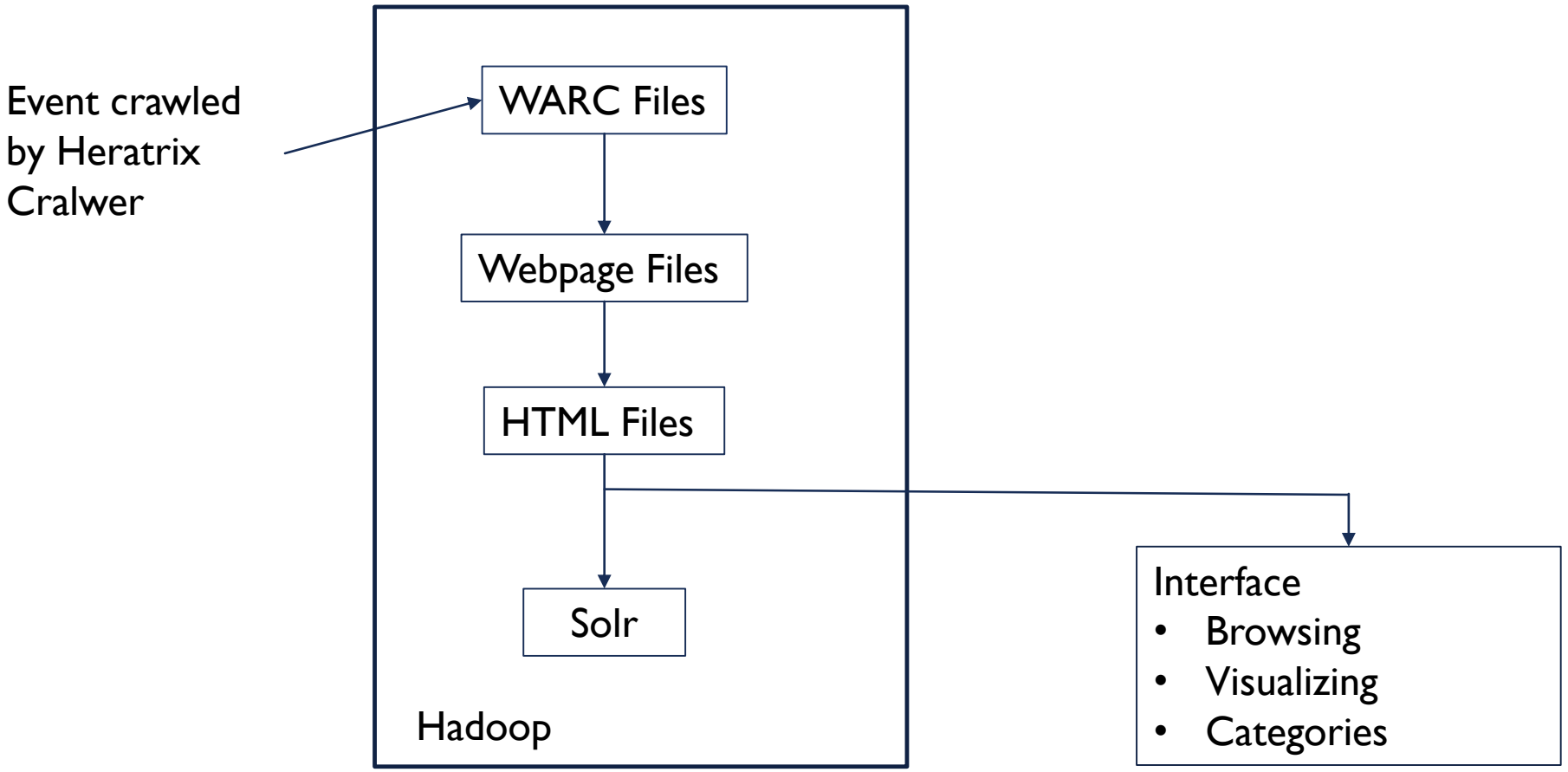
IDEAL PROJECT – (BORROWED FROM SUNSHIN LEE)



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
- Make the process reusable on other archives

PROJECT ARCHITECTURE



OUR ROLES

- .warc file extraction
- Filtering of HTML files
- Text extracting from HTML files
- Indexing information into Solr
- Map/Reduce Script for Hadoop

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
- Expand process of extracting .warc files to multiple files/directories
- Extracting text from HTML files
- Indexing information into Solr

EXTRACTING WARC FILE

- Integrated Hanzo Warc Tools (https://pypi.python.org/pypi/hanzo-warc-tools/0.2)
  - Only takes one warc file at a time
  - Unpacks warc file into HTTP folder and HTTPS folder
  - Creates text file to be used later
- Created script to allow for full directory to be unpacked

WARC EXTRACTION EXAMPLE OUTPUT

warc_file	warc_con_len	warc_uri_date	warc_subject_uri	uri_content_type	outfile
fn.warc.gz	3284	2013-04-21	www.vt.edu/robots.txt	text/plain	/Users/http/robots/txt
fn2.warc.gz	1023	2013-04-21	www.vt.edu/cs.html	text/html	/Users/https/cs.html
fn3.warc.gz	4983	2013-04-21	www.vt.edu/logo.png	image/jpg	/Users/http/logo.png

HTML EXTRACTION

- Find HTML documents based on uri_content_type column in text file
- Use the outfile column to locate where the file is an extract HTML file

INDEXING FILES INTO SOLR

- Text extracted from HTML files using BeautifulSoup4 (http://www.crummy.com/software/BeautifulSoup/)
- Indexed into Solr using solrpy (https://code.google.com/p/solrpy/)
- Use fields for:
  - id
  - content
  - collection_id
  - event
  - event_type
  - URL
  - wayback_URL

WORK REMAINING

- Work with client to integrate the process with Hadoop

QUESTIONS?