CS6604 Digital Libraries

IDEAL Webpages

Presented by
Ahmed Elbery, Mohammed Farghally
Project client
Mohammed Magdy
Virginia Tech, Blacksburg

3/5/2014
Overview

- A tremendous amount $\approx 10\text{TB}$ of data is available about a variety of events crawled from the web.
- It is required to make this big data accessible and searchable conveniently through the web.
- $\approx 10\text{TB}$ of .warc.
- Use only html files.
Tools required

- **Solr:**
- **Hadoop:**
- **Solarium**
- **PHP:**
- **Python/JAVA**
Big picture

Crawled Data → Hadoop → Solr → Index
Implementaion

Mohamed Seddik

Web Interface --> Search requests(AJAX) --> PHP Module --> Query --> Solarium --> Solr Server

Results

Response

Response (JSON or XML)

Ahmed Elbery

WARC Files --> Hadoop Uploader Module --> Extraction/Filtering Module

Map/Reduce

.html Files

Indexer Module

Index
THANK YOU

Mohammed Farghally
&
Ahmed Elbery

3/5/2014