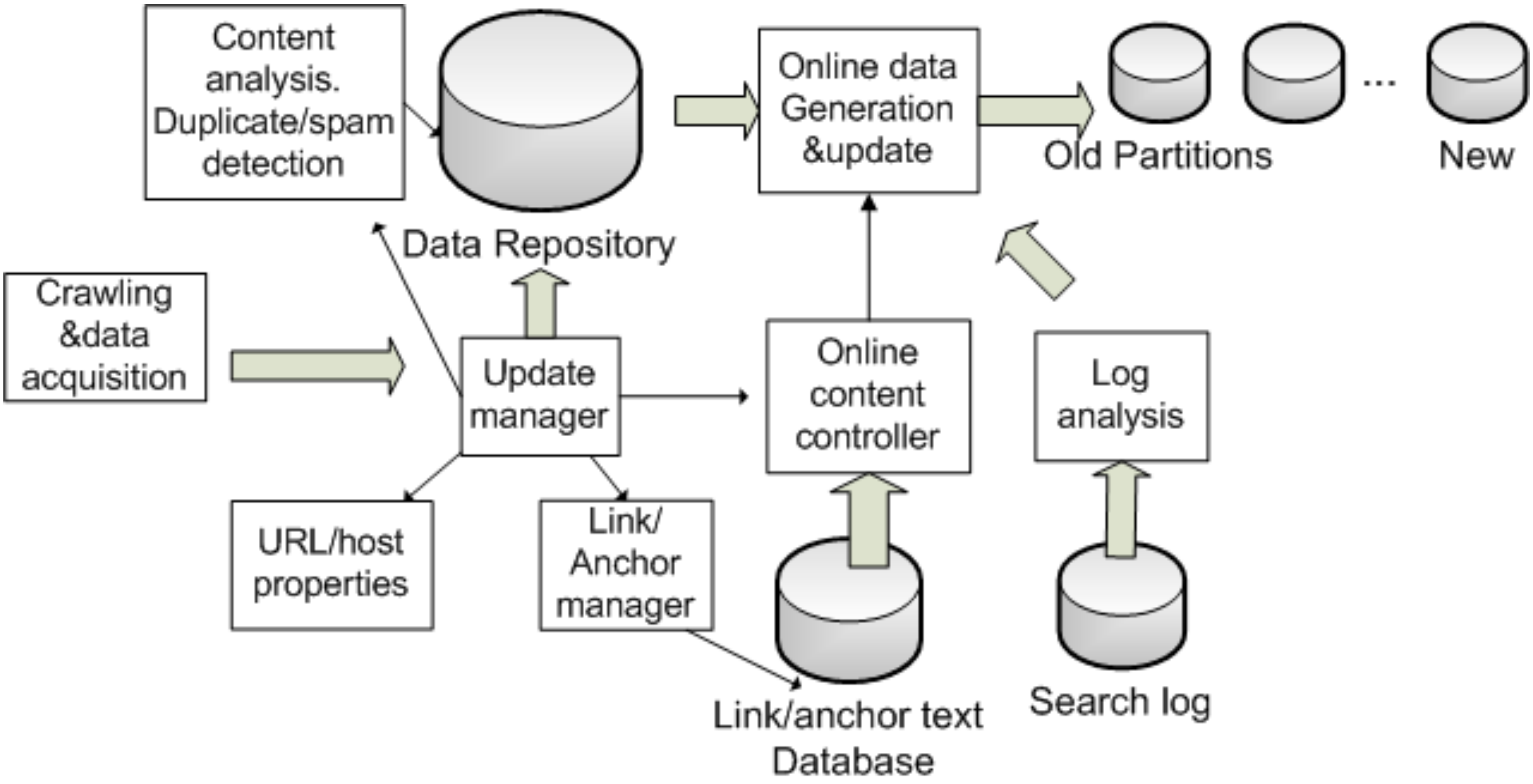
Offline Data Processing: Tasks and Infrastructure Support

T. Yang, UCSB 293S

Table of Content

- Offline incremental data processing: case study
- Example of content analysis
- System support

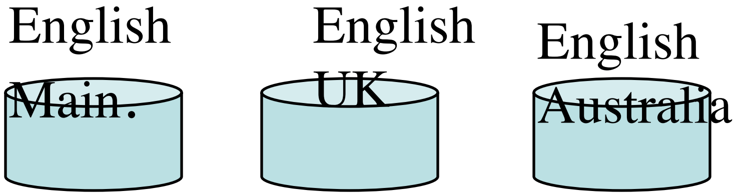
Offline Architecture for Ask.com Search


Content analysis. Duplicate/spam detection
Data Repository
Online data Generation &update
...
Old Partitions
New
Crawling &data acquisition
Update manager
Online content controller
Log analysis
URL/host properties
Link/ Anchor manager
Link/anchor text Database
Search log

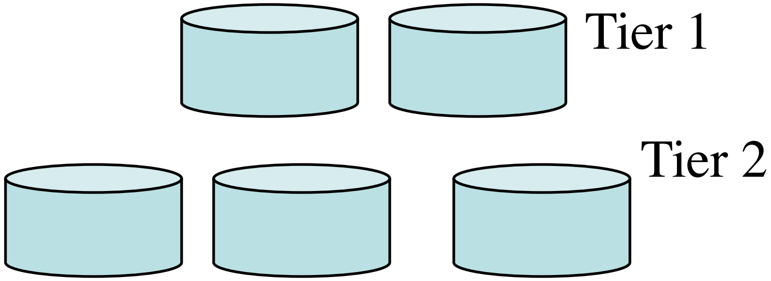
Content Management

- **Organize the vast amount of pages crawled to facilitate online search.**
  - Data preprocessing
  - Inverted index
  - Compression
  - Classify and partition data
- **Collect additional content and ranking signals.**
  - Link, anchor text, log data
- **Extract and structure content**
- **Duplicate detection**

Classifying and Partitioning data



- **Classify**
  - Content quality. Language/country etc
- **Partition**
  - Based on languages and countries. Geographical distribution based on data center locations
  - Partition based on quality
    - First tier --- high chance that users will access
      - Quality indicator
      - Click feedback
    - Second tier – lower chance

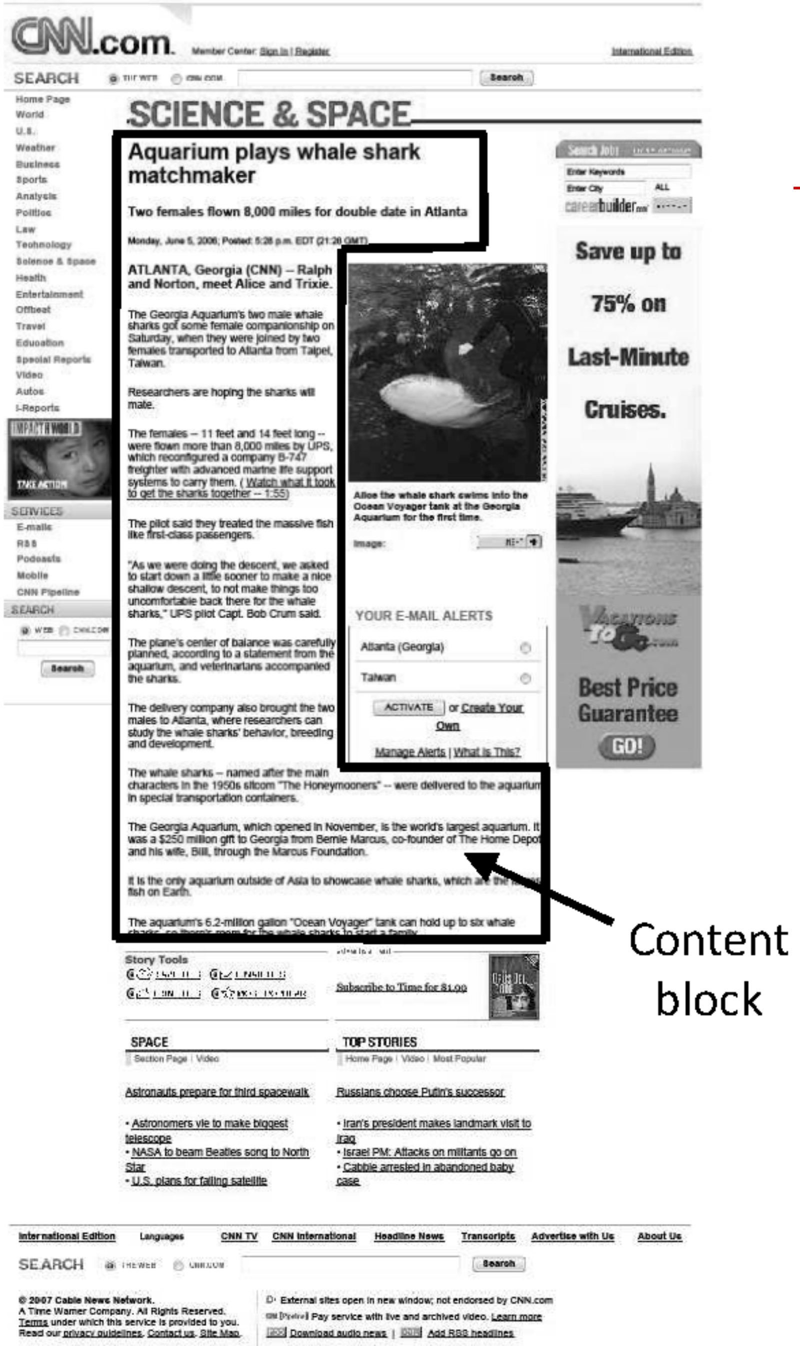


Examples of Context Extraction/Analysis

- **Identify key phrases that capture the meaning of this document.**
  - For example, title, section title, highlighted words.
    - HTML vs PDF
- **Identify parts of a document representing the meaning of this document.**
  - Many web pages contain a side-menu, which his less relevant to the main content of the documents
- **Capture page content through Javascript analysis.**
  - Page rendering and Javascript evaluation within a page

Example of Content Analysis

- **Detect content block related to the main content of a page**
  - Non-content text/link material is de-prioritized during indexing process



CNN.com Member Center: Sign In | Register International Edition

SEARCH ● THE WEB ○ CNN.COM Search

Home Page
World
U.S.
Weather
Business
Sports
Analysis
Politics
Law
Technology
Science & Space
Health
Entertainment
Offbeat
Travel
Education
Special Reports
Video
Autos
I-Reports

IMPACT YOUR WORLD TAKE ACTION

SERVICES
E-mails
RSS
Podcasts
Mobile
CNN Pipeline

SEARCH ● WEB ○ CNN.COM Search

SCIENCE & SPACE

Aquarium plays whale shark matchmaker

Two females flown 8,000 miles for double date in Atlanta

Monday, June 5, 2006; Posted: 5:28 p.m. EDT (21:28 GMT)

ATLANTA, Georgia (CNN) -- Ralph and Norton, meet Alice and Trixie.

The Georgia Aquarium's two male whale sharks got some female companionship on Saturday, when they were joined by two females transported to Atlanta from Taipei, Taiwan.

Researchers are hoping the sharks will mate.

The females -- 11 feet and 14 feet long -- were flown more than 8,000 miles by UPS, which reconfigured a company B-747 freighter with advanced marine life support systems to carry them. (Watch what it took to get the sharks together -- 1:55)

The pilot said they treated the massive fish like first-class passengers.

"As we were doing the descent, we asked to start down a little sooner to make a nice shallow descent, to not make things too uncomfortable back there for the whale sharks," UPS pilot Capt. Bob Crum said.

The plane's center of balance was carefully planned, according to a statement from the aquarium, and veterinarians accompanied the sharks.

The delivery company also brought the two males to Atlanta, where researchers can study the whale sharks' behavior, breeding and development.

The whale sharks -- named after the main characters in the 1950s sitcom "The Honeymooners" -- were delivered to the aquarium in special transportation containers.

The Georgia Aquarium, which opened in November, is the world's largest aquarium. It was a $250 million gift to Georgia from Bernie Marcus, co-founder of The Home Depot and his wife, Billi, through the Marcus Foundation.

It is the only aquarium outside of Asia to showcase whale sharks, which are the largest fish on Earth.

The aquarium's 6.2-million gallon "Ocean Voyager" tank can hold up to six whale sharks, so there's room for the whale sharks to start a family.

Alice the whale shark swims into the Ocean Voyager tank at the Georgia Aquarium for the first time.

Image:

YOUR E-MAIL ALERTS
Atlanta (Georgia) ○
Taiwan ○
ACTIVATE or Create Your Own
Manage Alerts | What Is This?

Search Jobs
Enter Keywords
Enter City ALL
careerbuilder.com

Save up to 75% on Last-Minute Cruises.
VacationsToGo.com
Best Price Guarantee
GO!

Story Tools

Subscribe to Time for $1.00

SPACE
Section Page | Video

Astronauts prepare for third spacewalk

- Astronomers vie to make biggest telescope
- NASA to beam Beatles song to North Star
- U.S. plans for falling satellite

TOP STORIES
Home Page | Video | Most Popular

Russians choose Putin's successor

- Iran's president makes landmark visit to Iraq
- Israel PM: Attacks on militants go on
- Cabbie arrested in abandoned baby case

International Edition Languages CNN TV CNN International Headline News Transcripts Advertise with Us About Us

SEARCH ● THE WEB ○ CNN.COM Search

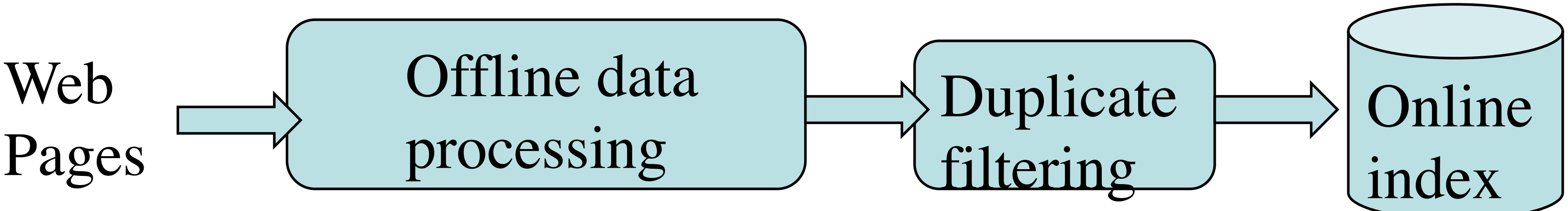
© 2007 Cable News Network.
A Time Warner Company. All Rights Reserved.
Terms under which this service is provided to you.
Read our privacy guidelines. Contact us. Site Map.

External sites open in new window; not endorsed by CNN.com
Pay service with live and archived video. Learn more
Download audio news | Add RSS headlines

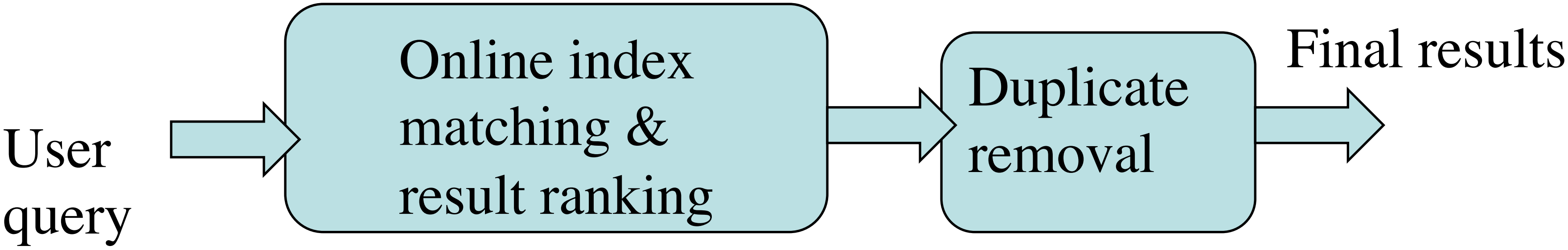
Content block

Redundant Content Removal in Search Engines

- **Over 1/3 of Web pages crawled are near duplicates**
- **When to remove near duplicates?**
  - Offline removal

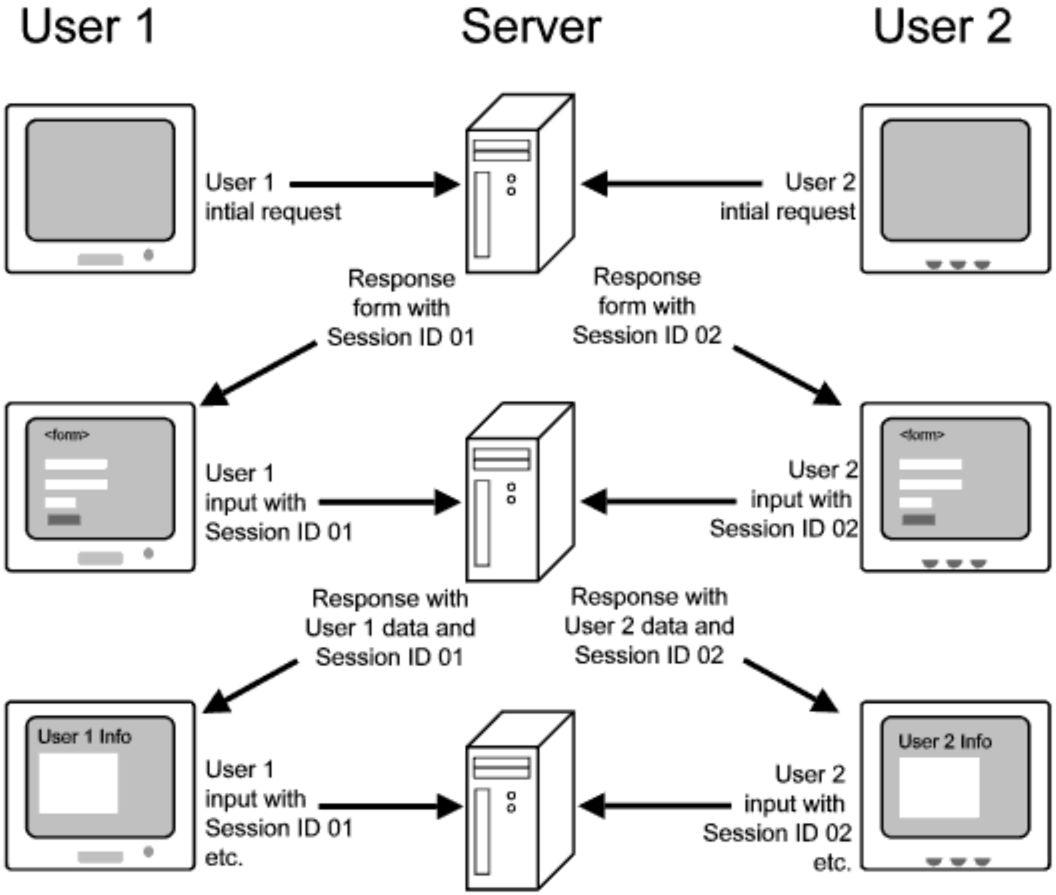


  - Online removal with query-based duplicate removal



Why there are so many duplicates?

- Same content, different URLs, often with different session IDs.
- Crawling time difference

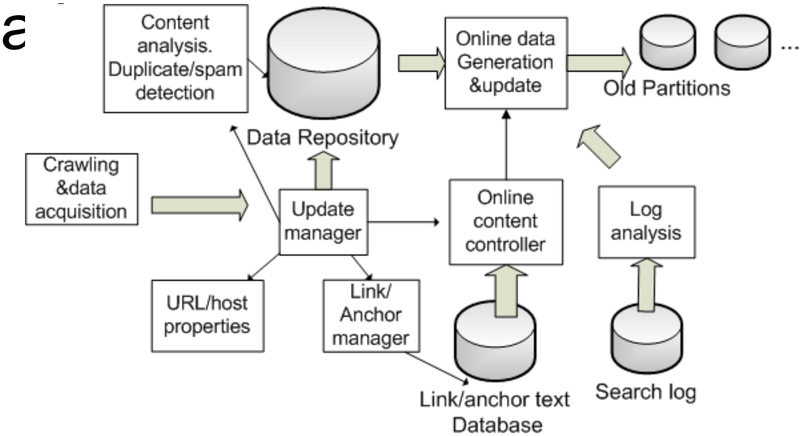

User 1
Server
User 2
User 1 intial request
User 2 intial request
Response form with Session ID 01
Response form with Session ID 02
<form>
<form>
User 1 input with Session ID 01
User 2 input with Session ID 02
Response with User 1 data and Session ID 01
Response with User 2 data and Session ID 02
User 1 Info
User 2 Info
User 1 input with Session ID 01 etc.
User 2 input with Session ID 02 etc.

Tradeoff of online vs. offline removal

	Online-dominating approach	**Offline-dominating approach**
Impact to offline	High precision Low recall Remove fewer duplicates	High precision High recall Remove most of duplicates Higher offline burden
Impact to online	More burden to online deduplication	Less burden to online deduplication
Impact to overall cost	Higher serving cost	Lower serving cost

Software Infrastructure Support at Ask.com

- **Programming support (multi-threading/exception Handling, Hadoop MapReduce)**
- **Data stores for managing billions of objects**
  - Distributed hash tables, queues etc
- **Communication and data exchange among machines/services**
- **Execution environment**
  - Controllable (stop, pause, restart).
  - Service registration and invoca
  - service monitoring
  - Logging and test framework.


Content analysis. Duplicate/spam detection
Data Repository
Online data Generation &update
Old Partitions
...
Crawling &data acquisition
Update manager
Online content controller
Log analysis
URL/host properties
Link/ Anchor manager
Link/anchor text Database
Search log

Requirements for Data Repository Support in Offline Systems

- **Update**
  - handling large volumes of modified documents
  - adding new content
- **Random access**
  - request the content of a document based on its URL
- **Compression and large files**
  - reducing storage requirements and efficient access
- **Scan**
  - Scan documents for text mining.

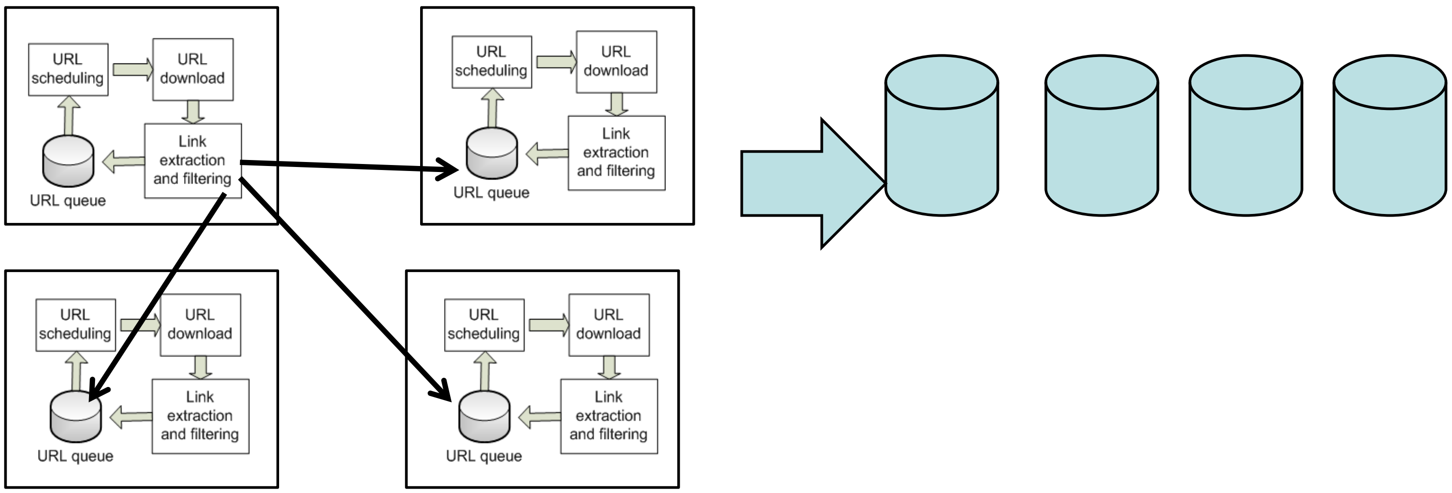
Options for Key-value Data Stores

- **Support: append or put. get operations**
- **Bigtable at Google**
- **Dynamo at Amazon**
- **Open source software**

	Technology	Language Platform	Users/ sponsors
Apache Cassandra	Bigtable Dynamo	Java/Hadoop	Apache
Hypertable	Bigtable	C++/Hadoop	Baidu
Hbase	Bigtable	Java/Hadoop	Apache
LevelDB	Bigtable	C++	Google
MongoDB		C++	

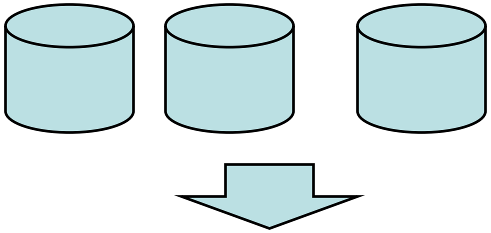
Sample Requirements for Applications: Data repository for crawling

- **Common data operations**
  - Update: Mainly append operations every day.
  - Content read:
    - Typically scan and then transfer data to another cluster
    - Sometime: random access individual pages for inspection



Sample Requirements for periodic data reclassification

- **Data repository hosting a large page collection with periodical page re-classification**
  - Update: Append only operations for raw data
    - Update → meta data modification periodically for selected pages (random access).
  - Read: Scan only operations for raw data processing.
    - Random read sometime for a small number of pages.



Data repository

MapReduce for classification