Crawlware - Seravia's Deep Web Crawling System

Presented by

邹志乐

robin@seravia.com

敬宓

jingmi@seravia.com

Agenda

- What is Crawlware?
- Crawlware Architecture
- Job Model
- Payload Generation and Scheduling
- Rate Control
- Auto Deduplication
- Crawler testing with Sinatra
- Some problems we encountered & TODOs

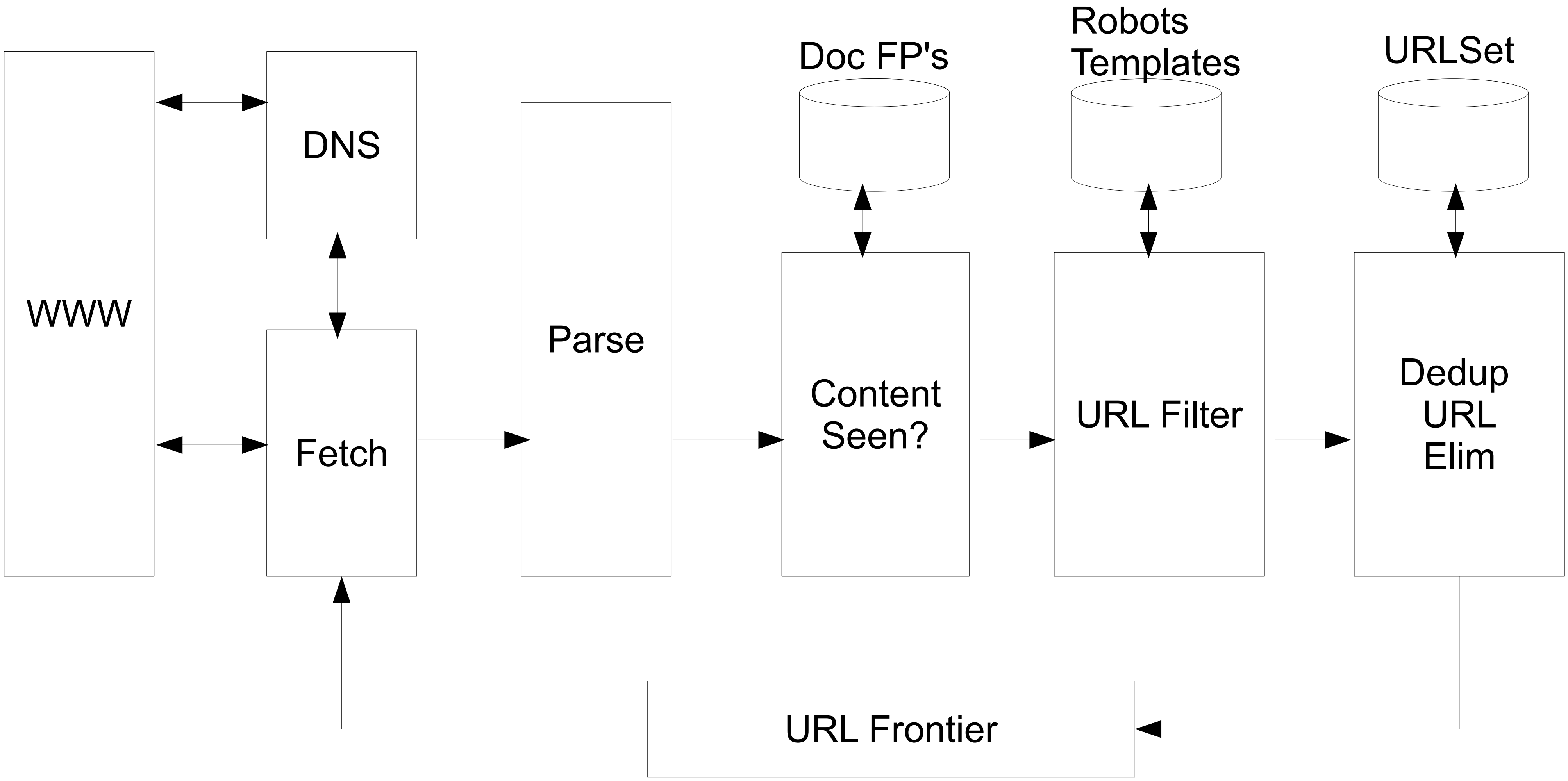
What is Crawlware?

Crawlware is a distributed deep web crawling system, which enables scalable and friendly crawls of the data that must be retrieved with complex queries.

- **Distributed**: Execute cross multiple machines
- **Scalable**: Scale up by adding extra machines and bandwidth.
- **Efficiency**: Efficient use of various system resources
- **Extensible**: Be extensible for new data formats, protocols, etc
- **Freshness**: Be able to capture data changes
- **Continuous**: Continuous crawling without administrators' operation.
- **Generic**: Each crawling worker can crawl any given sites
- **Parallelization**: Crawl all websites in parellel
- **Anti-blocking**: Precise rate control

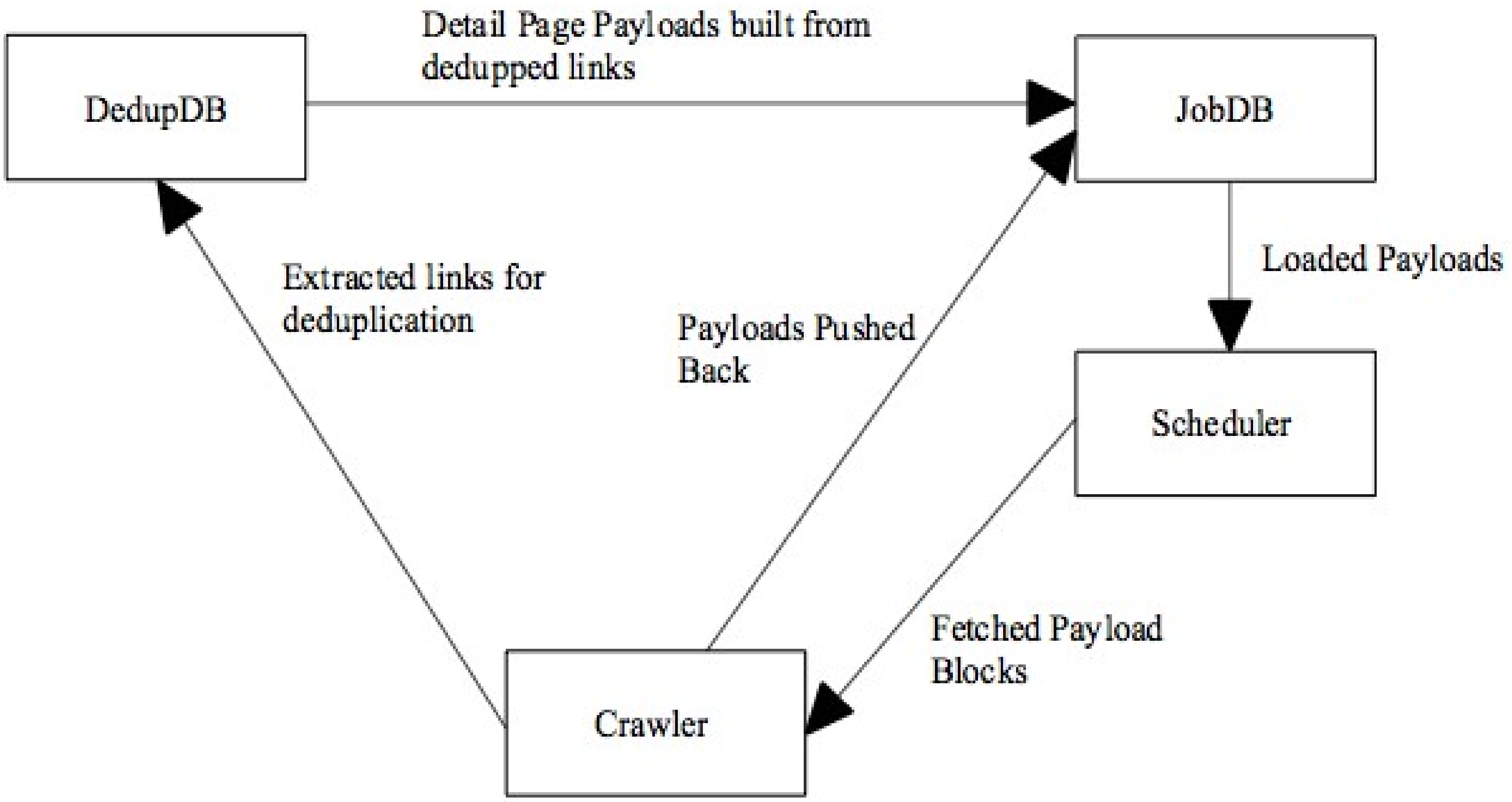
A General Crawler Architecture

From <<Introduction to Information Retrieval>>

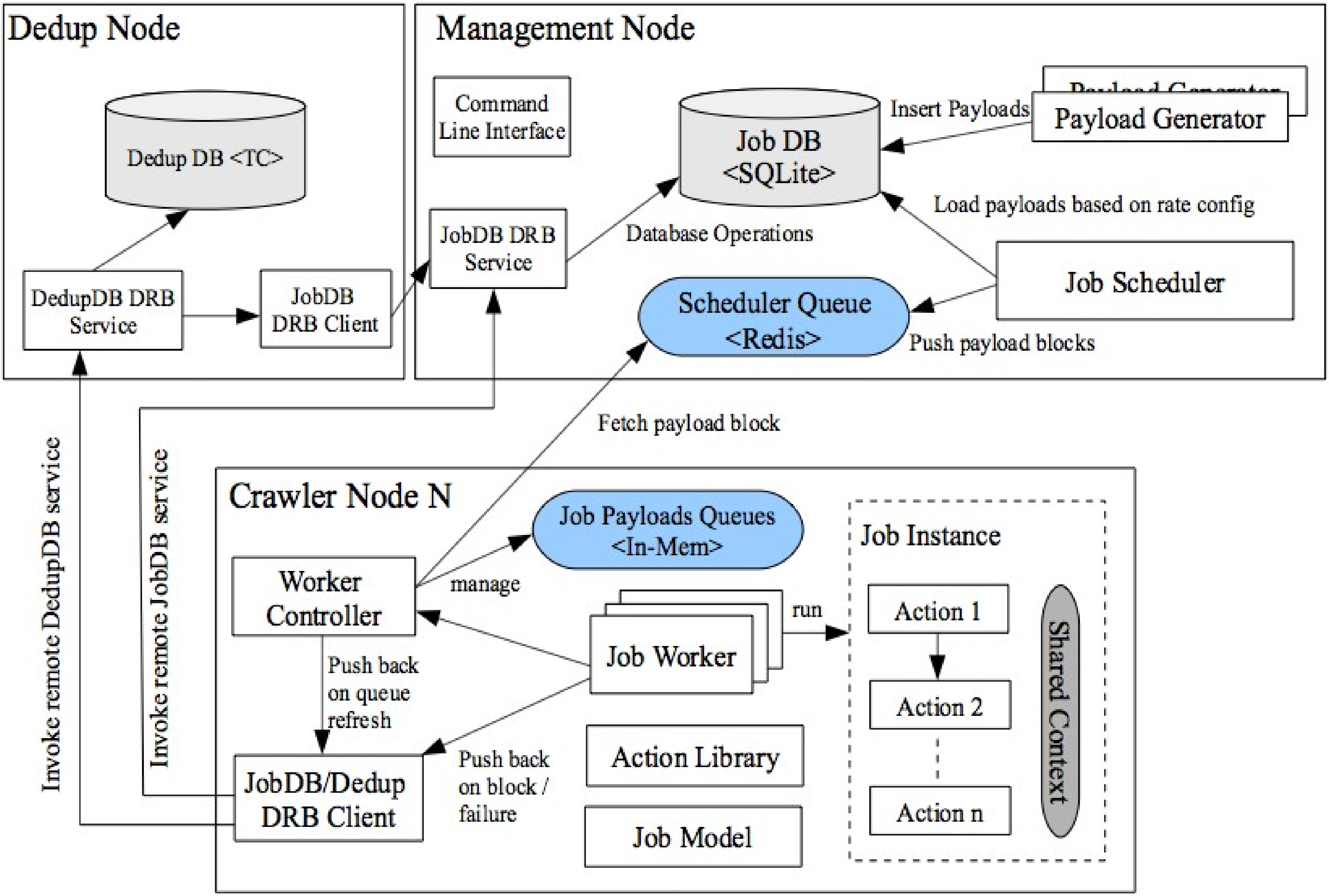


Crawlware Architecture

High Level Working Flows



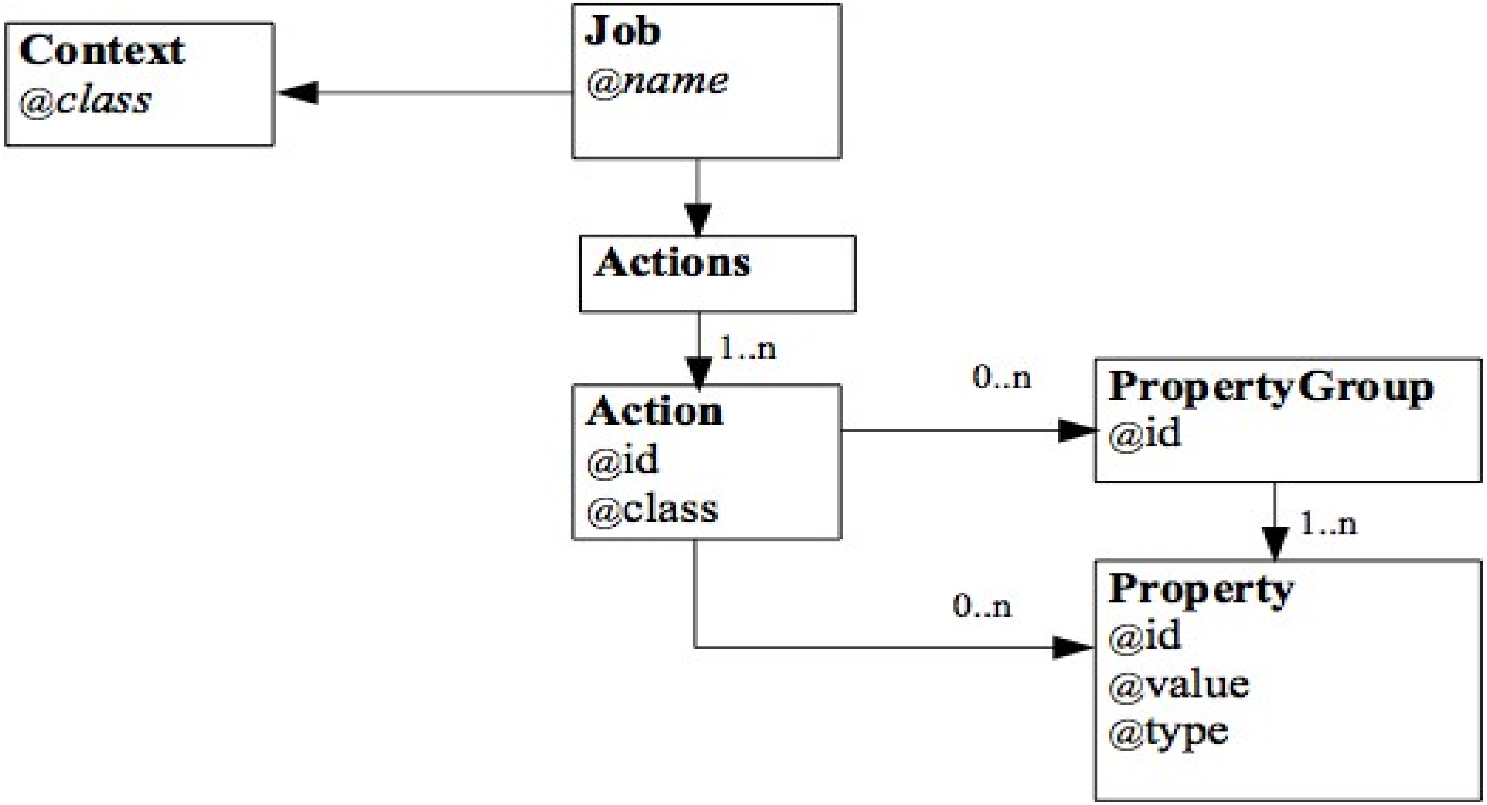
Crawlware Architecture



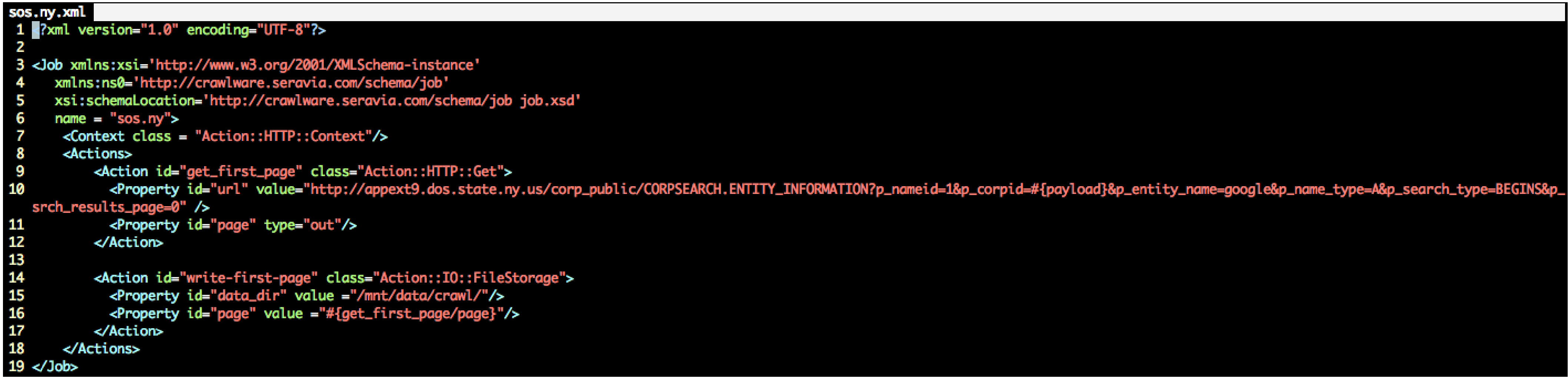
Job Model

- XML syntax
- An assembly of reusable actions
- A context shared at runtime
- Job & Actions customized through properties

- HTTP Get, Post, Next Page
- Page Extractor, Link Extractor
- File Storage
- Assignment
- Code Snippet



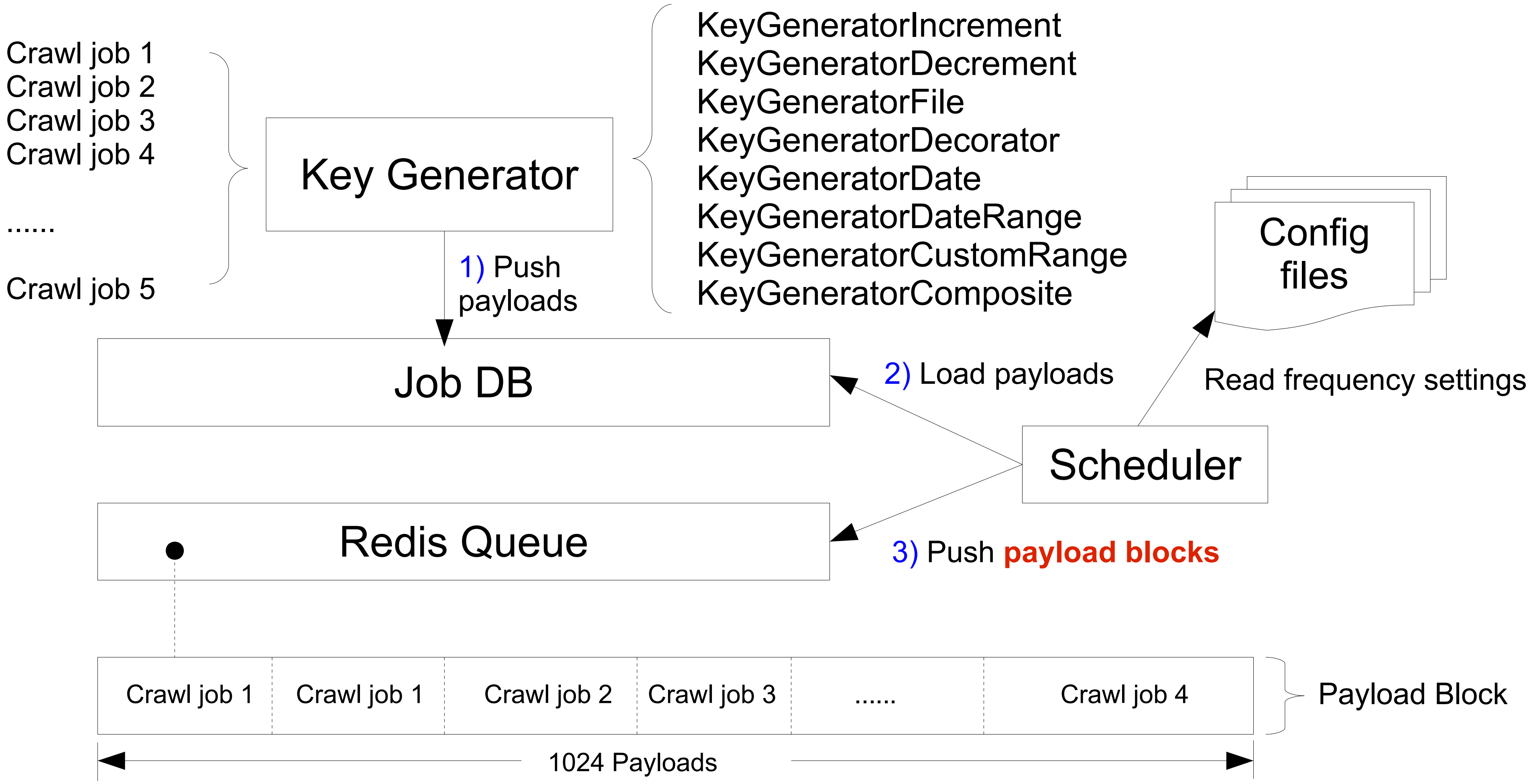
Job Model Sample



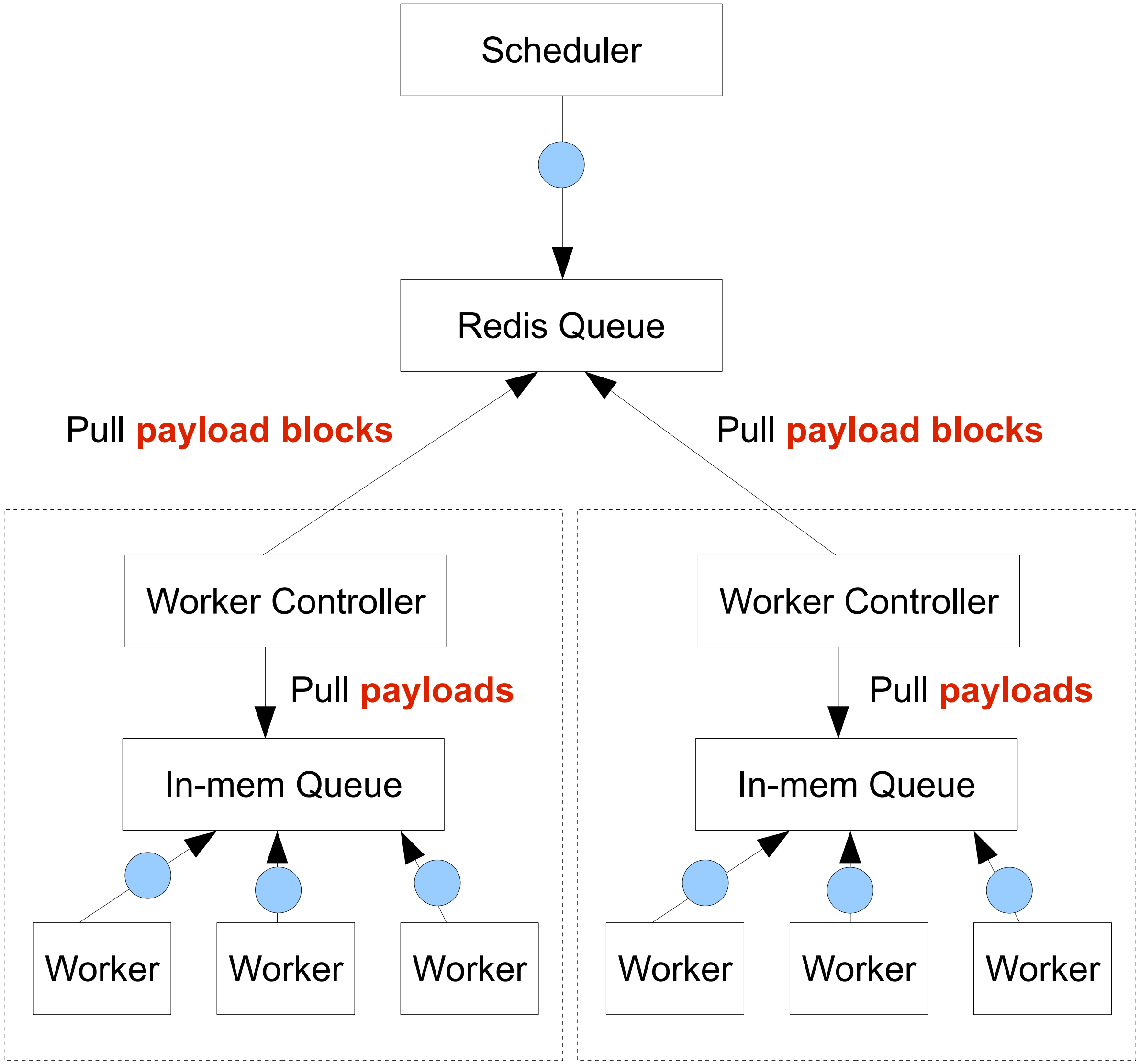
sos.ny.xml

```
<?xml version="1.0" encoding="UTF-8"?>

<Job xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
   xmlns:ns0='http://crawlware.seravia.com/schema/job'
   xsi:schemaLocation='http://crawlware.seravia.com/schema/job job.xsd'
   name = "sos.ny">
    <Context class = "Action::HTTP::Context"/>
    <Actions>
        <Action id="get_first_page" class="Action::HTTP::Get">
          <Property id="url" value="http://appext9.dos.state.ny.us/corp_public/CORPSEARCH.ENTITY_INFORMATION?p_nameid=1&p_corpid=#{payload}&p_entity_name=google&p_name_type=A&p_search_type=BEGINS&p_srch_results_page=0" />
          <Property id="page" type="out"/>
        </Action>

        <Action id="write-first-page" class="Action::IO::FileStorage">
          <Property id="data_dir" value ="/mnt/data/crawl/"/>
          <Property id="page" value ="#{get_first_page/page}"/>
        </Action>
    </Actions>
</Job>
```

Payload Generation & Scheduling

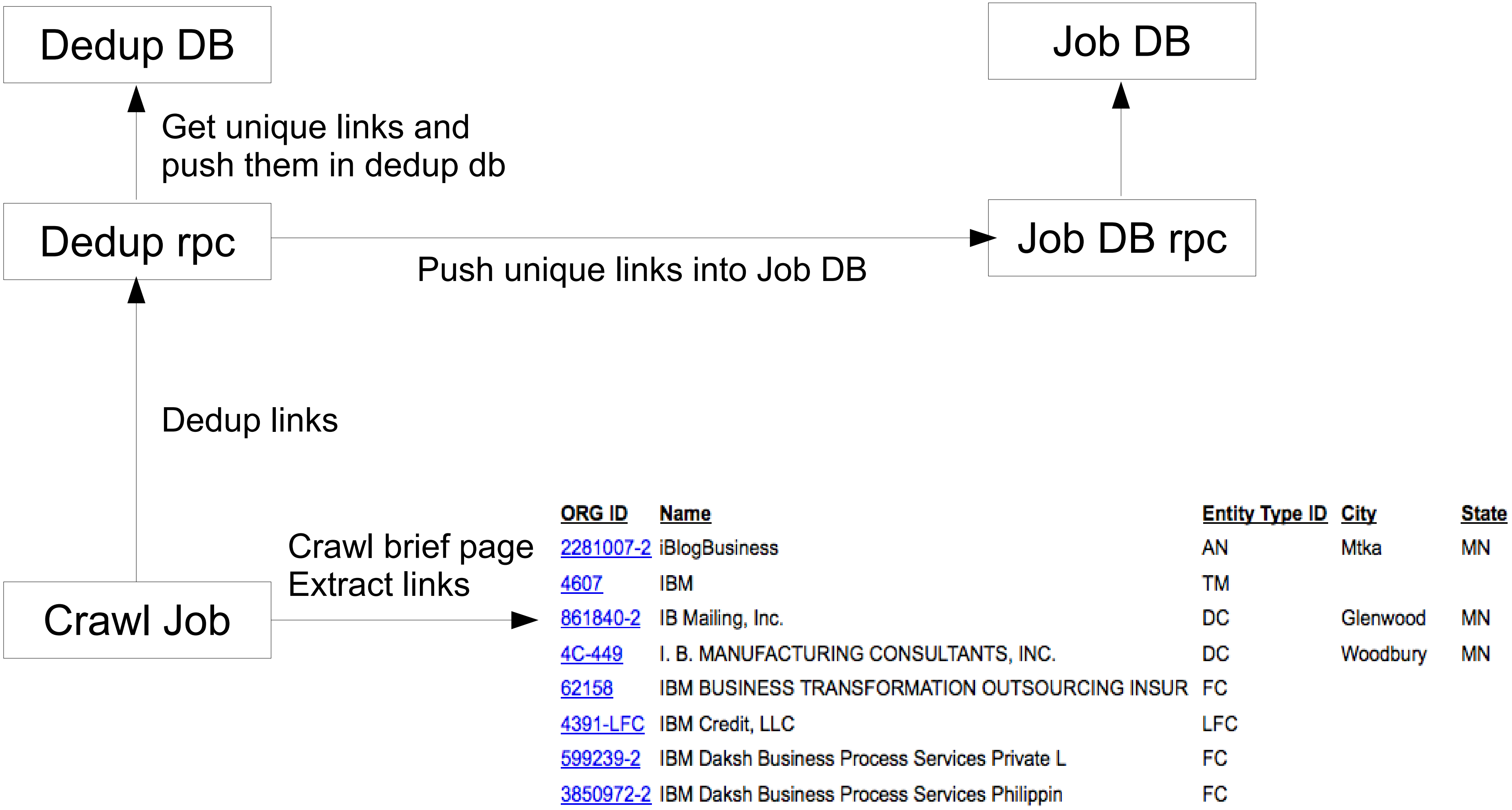


Rate Control



- Site frequency configuration
- A given site's payloads amount in payload block is determined by the crawling frequency.
- Scheduler controls the crawling rate of the entire system (N crawler nodes/IPs)
- Worker Controller controls the crawling rate of a single node/IP

Auto Deduplication



Dedup DB

Job DB

Get unique links and
push them in dedup db

Dedup rpc

Job DB rpc

Push unique links into Job DB

Dedup links

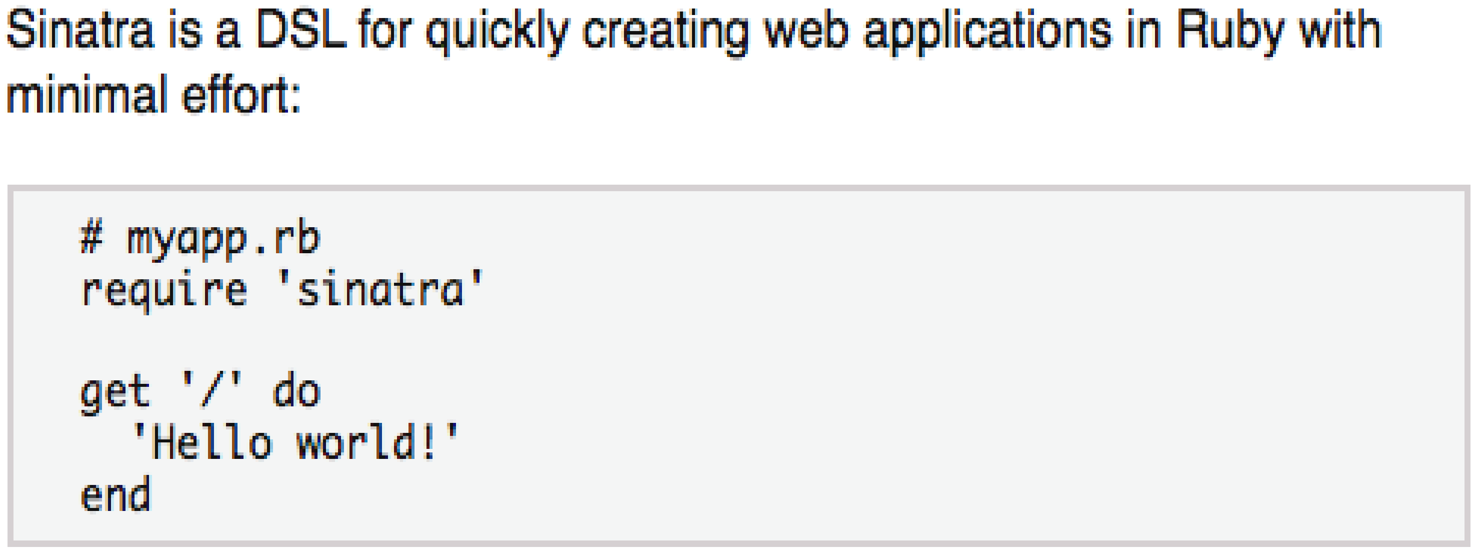
Crawl brief page
Extract links

Crawl Job

ORG ID	Name	Entity Type ID	City	State
2281007-2	iBlogBusiness	AN	Mtka	MN
4607	IBM	TM		
861840-2	IB Mailing, Inc.	DC	Glenwood	MN
4C-449	I. B. MANUFACTURING CONSULTANTS, INC.	DC	Woodbury	MN
62158	IBM BUSINESS TRANSFORMATION OUTSOURCING INSUR	FC		
4391-LFC	IBM Credit, LLC	LFC		
599239-2	IBM Daksh Business Process Services Private L	FC		
3850972-2	IBM Daksh Business Process Services Philippin	FC		

Crawler Testing with Sinatra

- What is Sinatra



Sinatra is a DSL for quickly creating web applications in Ruby with minimal effort:

```
# myapp.rb
require 'sinatra'

get '/' do
  'Hello world!'
end
```

- Crawler Testing
  - Simulate various crawling actions via http, such as Get, Post, NextPage.
  - Simulate job profiles

Encountered Problems & TODOs

- Changing site load/performance
  - Monitoring
  - Dynamic rate switch based on time zone
- Page Correctness
  - Page tracker – continuous errors or identical pages
- Data Freshness
  - Scheduled updates
  - Crawl delta for ID or date range payloads
  - Recrawl for keyword payloads
- Javascript

Thank You

Please contact jobs@seravia.com for job opportunities.