

Building Babel

Large Scale Data Collection in the Cloud

Ian Wesley-Smith

iwsmith@uw.edu



Information School
UNIVERSITY of WASHINGTON



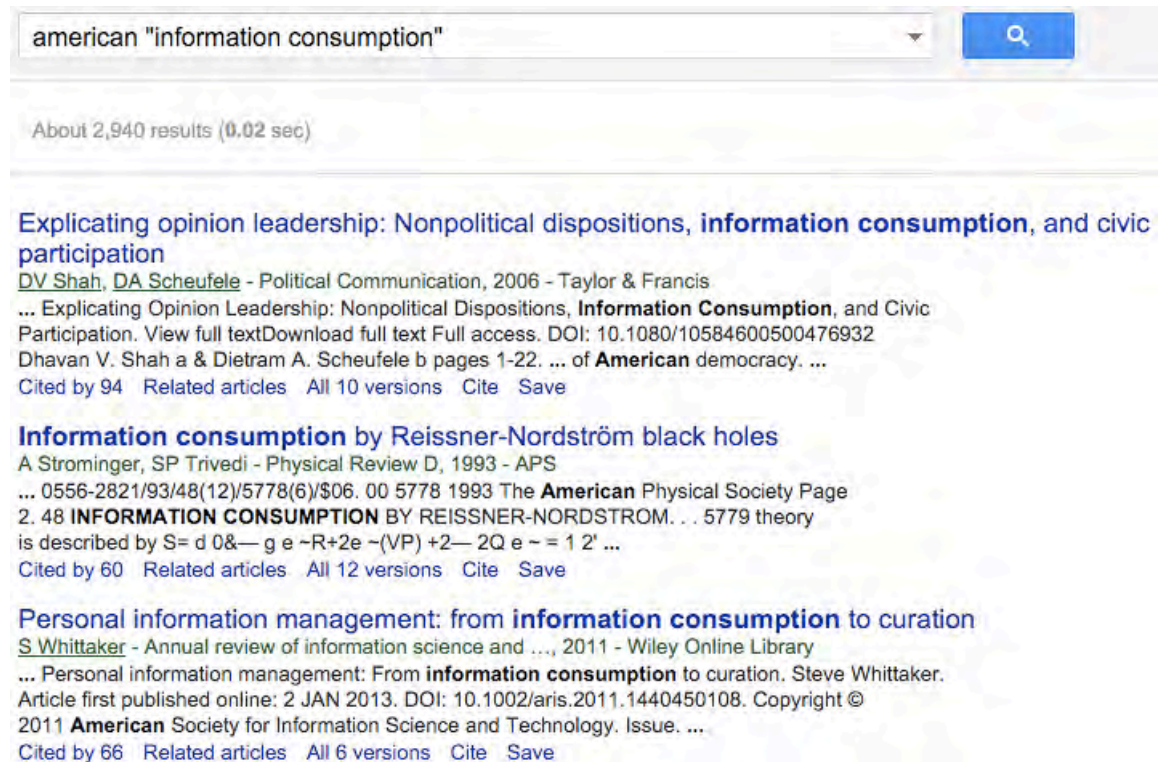
Scholarly Article Recommendation

- Information Overload
 - 50m – 150m articles in existence



Google Scholar

- Recommendation vs Search
 - Serendipity
- Homonymy
- Synonymy



The screenshot shows a Google Scholar search interface. The search bar contains the text "american 'information consumption'" and a search button with a magnifying glass icon. Below the search bar, it indicates "About 2,940 results (0.02 sec)". Three search results are displayed, each with a title, author information, and publication details.

american "information consumption"

About 2,940 results (0.02 sec)

Explicating opinion leadership: Nonpolitical dispositions, information consumption, and civic participation
DV Shah, DA Scheufele - Political Communication, 2006 - Taylor & Francis
... Explicating Opinion Leadership: Nonpolitical Dispositions, **Information Consumption**, and Civic Participation. View full textDownload full text Full access. DOI: 10.1080/10584600500476932
Dhavan V. Shah a & Dietram A. Scheufele b pages 1-22. ... of **American** democracy. ...
Cited by 94 Related articles All 10 versions Cite Save

Information consumption by Reissner-Nordström black holes
A Strominger, SP Trivedi - Physical Review D, 1993 - APS
... 0556-2821/93/48(12)/5778(6)/\$06.00 5778 1993 The **American** Physical Society Page
2. 48 **INFORMATION CONSUMPTION** BY REISSNER-NORDSTROM. ... 5779 theory
is described by $S = d \int g e^{-R+2e} \sim (VP) + 2 \int Q e \sim 1 2' \dots$
Cited by 60 Related articles All 12 versions Cite Save

Personal information management: from information consumption to curation
S Whittaker - Annual review of information science and ..., 2011 - Wiley Online Library
... Personal information management: From **information consumption** to curation. Steve Whittaker.
Article first published online: 2 JAN 2013. DOI: 10.1002/aris.2011.1440450108. Copyright ©
2011 **American** Society for Information Science and Technology. Issue. ...
Cited by 66 Related articles All 6 versions Cite Save



Netflix/Spotify/Amazon

- User ratings (explicit, implicit)
- Density
 - # user-item interactions \gg # items
- Netflix Competition (2006)¹
 - 100m ratings
 - 480k users
 - 17k movies

1: <http://www.netflixprize.com/community/viewtopic.php?id=68>



Barriers to Research

- Hard to get datasets
- Difficult to measure effectiveness
 - Judges
 - Citation prediction



Enter Babel

- Provide access to private data sets
- Provide scholarly article recommendations, freely to anyone
 - Feedback data in return
- Evaluate recommenders using usage data
 - With enough traffic could be very fast



Audience

- Publishers
 - Offload expensive research into recommender systems to academia
 - Better recommendations drive more traffic/purchases
- Tool Developers
- Researchers



Requirements

- Fast
- Reliable
- Scalable (lots of data!)
- Easy to use
- Cheap



REST API

```
curl http://babel-us-east-
1.eigenfactor.org/recommendation/aminer/12345
{
  "transaction_id": "46bb84190e9ddfd17700bfafb500ab3c",
  "results": [
    {
      "paper_id": "672",
      "publisher": "aminer"
    },
    {
      "paper_id": "11274",
      "publisher": "aminer"
    }
  ]
}
```



http://babel.eigenfactor.org

Babel **Home** About API

Eigenfactor Search

AMiner arXiv DBLP JSTOR MAS PLOS PubMed

Search Results

Get Related The Eigenfactor Metrics. WisemanMA **reviews meta-analyses**

Get Related Assessing Citations With The Eigenfactor™ Metrics. WestJ **reviews meta-analyses**

Get Related The Most Influential Journals: Impact Factor And Eigenfactor. FershtA **impact factor**

Get Related Impact Factor, Eigenfactor And Article Influence AscasoFJ **reviews meta-analyses**

Get Related Big Macs And Eigenfactor Scores: Don't Let Correlation Coefficients Fool You J West 2009 **behavior small-world**

Get Related The Relation Between Eigenfactor, Audience Factor, And Influence Weight L Waltman 1910 **behavior small-world**

Get Related Eigenfactor: Does The Principle Of Repeated Improvement Result In Better Estimates Than Raw Citat... P Davis 2007 **the anatomy**

Get Related The Relation Between Eigenfactor, Audience Factor, And Influence Weight. L Waltman 2009 **the anatomy**

Get Related The Relation Between Eigenfactor, Audience Factor, And Influence Weight L Waltman 2009 **the anatomy**

Get Related Comparison Between Impact Factor, SCImago Journal Rank Indicator And Eigenfactor Score Of Nuclear... S A **reviews meta-analyses**

« Previous **1** 2 Next »

Papers related to

The Eigenfactor Metrics. WisemanMA **reviews meta-analyses**

Get Related Evaluating "Payback" On Biomedical Research From Papers Cited In Clinical Guidelines: An Applied ... FawcettG **reviews meta-analyses**

Get Related Searching For Intellectual Turning Points: Progressive Knowledge Domain Visualization. ChenC **reviews meta-analyses**

Get Related Evaluating Research And Impact: A Bibliometric Analysis Of Research By The NIH/NIAID HIV/AIDS Cli... S Johan 2011 **reviews meta-analyses**

Get Related Translation Of Research Into Practice: Why We Can't "Just Do It". SeifertCM **reviews meta-analyses**



Browser Plugins



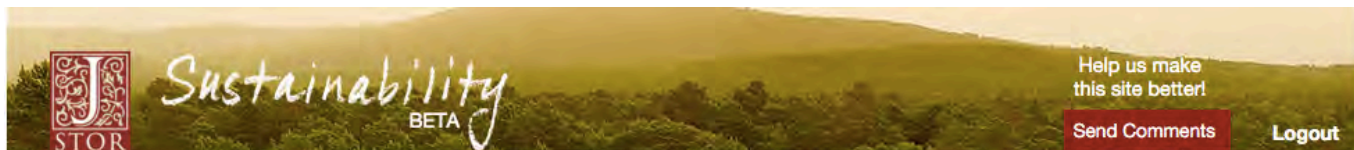
The screenshot shows the Firefox browser window with the address bar containing the URL `https://scholar.google.com/scholar?q=Eigenfactor&btnG=&hl=en&as_sdt=0%2C48`. The search results page displays the Google logo, the search term 'Eigenfactor', and the number of results: 'About 38,900 results (0.05 sec)'. The results are categorized into 'Articles', 'Case law', and 'My library'. The top article is 'The Eigenfactor™ metrics' by CT Bergstrom and JD West, published in The Journal of ... in 2008. Other articles include 'Networks of scientific papers' by P Yu and H Van de Sompel, and 'The most influential journals: Impact Factor and Eigenfactor' by A Fersht. The left sidebar contains filters for 'Any time' (Since 2015, 2014, 2011, Custom range...), 'Sort by relevance', 'Sort by date', and checkboxes for 'include patents' and 'include citations'. A 'Create alert' button is also visible.



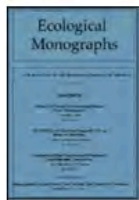
The screenshot shows the Chrome browser window with the address bar containing the URL `https://scholar.google.com/scholar?hl=en&q=Eigenfactor&btnG=&as_sdt=1%2C48&as_sdtp=`. The search results page displays the Google logo, the search term 'Eigenfactor', and the number of results: 'About 38,900 results (0.02 sec)'. The results are categorized into 'Articles', 'Case law', and 'My library'. The top article is 'The Eigenfactor™ metrics' by CT Bergstrom and JD West, published in The Journal of ... in 2008. Other articles include 'Networks of scientific papers' by P Yu and H Van de Sompel, and 'The most influential journals: Impact Factor and Eigenfactor' by A Fersht. The left sidebar contains filters for 'Any time' (Since 2015, 2014, 2011, Custom range...), 'Sort by relevance', 'Sort by date', and checkboxes for 'include patents' and 'include citations'. A 'Create alert' button is also visible.



http://labs.jstor.org/sustainability/



Search 



Cross-Scale Morphology, Geometry, and Dynamics of Ecosystems

C. S. Holling
Ecological Monographs
Ecological Society of America

Vol. 62, No. 4 (Dec., 1992), pp. 447-502

Stable URL: www.jstor.org/stable/2937313



Abstract

This paper tests the proposition that a small set of plant, animal, and abiotic processes structure **ecosystems** across scales in time and space. Earlier studies have suggested that these key structuring processes establish a small number of dominant temporal frequencies that entrain other processes. These frequencies often differ from each other by at least an order of ...

[▶ More](#)

Topics

Boreal forests , Animal physiology , Prairies , Mammals , Terrestrial ecosystems , Landscapes , Birds , **Ecosystems** , Animals , Species

Background Reading

Command and Control and the Pathology of Natural Resource Management
C. S. Holling, Gary K. Meffe
Conservation Biology, Vol. 10, No. 2 (Apr., 1996), pp. 328-337

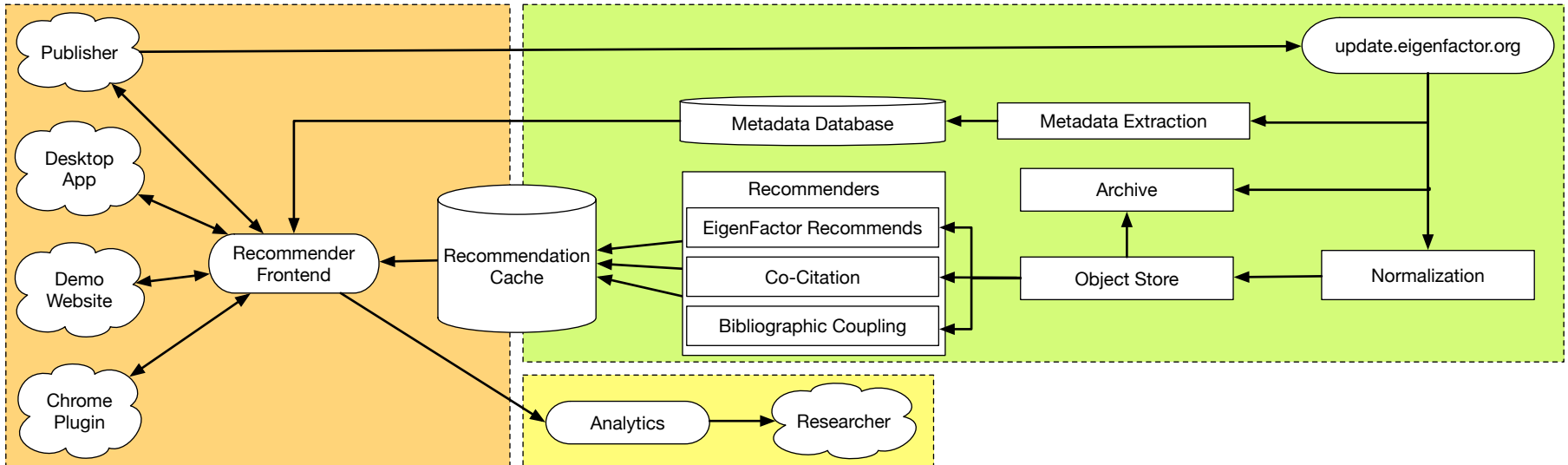
Management of Eutrophication for Lakes Subject to Potentially Irreversible Change
S. R. Carpenter, D. Ludwig, W. A. Brock
Ecological Applications, Vol. 9, No. 3 (Aug., 1999), pp. 751-771

[More Recommendations ▶](#)

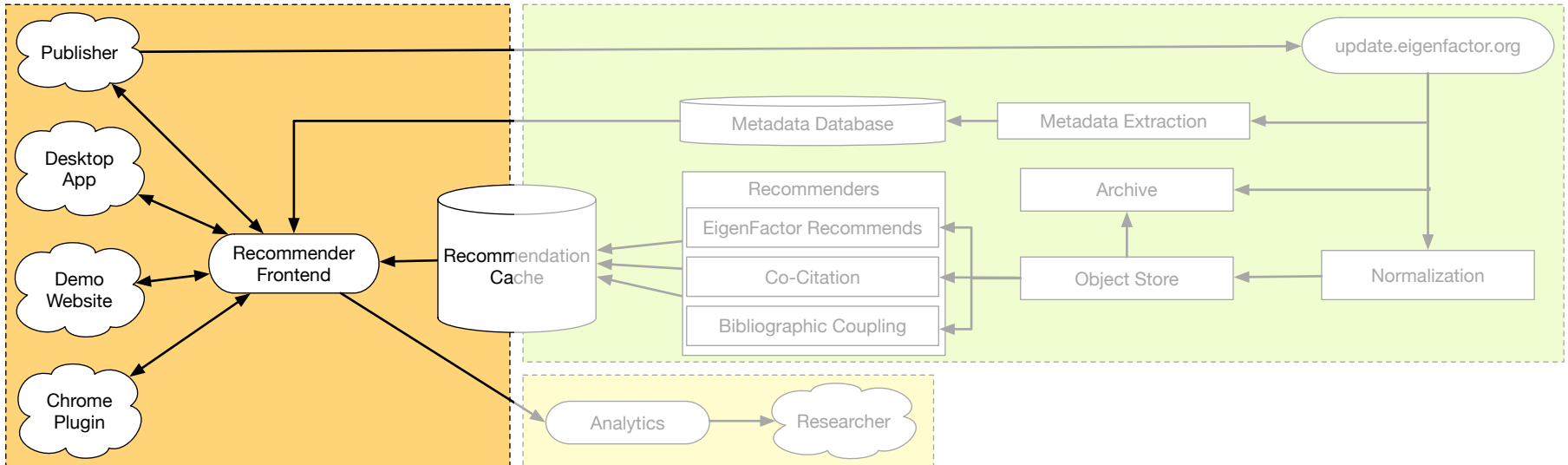
[< Previous Item](#) | [Next Item >](#)



Babel Architecture



Frontend



Frontend

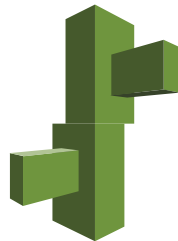
Application



Package



Deploy



AWS Elastic Bean Stalk



Frontend

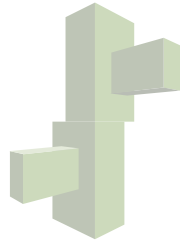
Application



Package



Deploy



AWS Elastic Bean Stalk





Swagger UI interface showing the definition of the `GET /recommendation/{publisher}/{paper_id}` endpoint.

```
268     description: No metadata found
269 |   '/recommendation/{publisher}/{paper_id}':
270 |     get:
271 |       description: Generates recommendations for `paper_id`
272 |       operationId: application.get_recommendation
273 |       parameters:
274 |         - $ref: '#/parameters/paper_id'
275 |         - $ref: '#/parameters/publisher'
276 |         - $ref: '#/parameters/client_id'
277 |         - default: 5
278 |           description: Maximum number of recommendations to return
279 |           in: query
280 |           maximum: 10
281 |           minimum: 1
282 |           name: limit
283 |           required: false
284 |           type: integer
285 |         - description: Algorithm to generate recommendations with. If not provided a
286 |           random algorithm will be used.
287 |           enum:
288 |             - ef_expert
289 |             - ef_classic
290 |           in: query
291 |           name: algorithm
292 |           required: false
293 |           type: string
294 |       produces:
295 |         - application/json
296 |       responses:
297 |         '200':
298 |           description: Successful response
299 |           schema:
300 |             $ref: '#/definitions/Recommendations'
301 | /search:
302 |   get:
303 |     description: 'Searches metadata for known papers, authors or labels'
304 |     operationId: application.search
305 |     parameters:
306 |       - description: Search query
307 |         in: query
308 |         name: q
309 |         required: true
310 |         type: string
311 |       - collectionFormat: multi
312 |         description: Publishers to restrict search to.
313 |         in: query
```

GET /recommendation/{publisher}/{paper_id}

Description
Generates recommendations for `paper_id`

Parameters

Name	Located in	Description	Required	Schema
<code>paper_id</code>	path	Publisher assigned identifier of a paper	Yes	↔ string
<code>publisher</code>	path	Publisher to perform this operation on	Yes	↔ string
<code>client_id</code>	query	Identifier provided by the platform to clients to track client usage.	No	↔ string
<code>limit</code>	query	Maximum number of recommendations to return	No	↔ integer
<code>algorithm</code>	query	Algorithm to generate recommendations with. If not provided a random algorithm will be used.	No	↔ string

Responses

Code	Description	Schema
200	Successful response	↔ Recommendations { results: ▶ [] transaction_id: ▶ string }



Swagger UI

GET /recommendation/{publisher}/{paper_id}

Implementation Notes
Generates recommendations for paper_id

Response Class (Status 200)
Model Model Schema

```
{
  "results": [
    {
      "paper_id": "string",
      "publisher": "string"
    }
  ],
  "transaction_id": "string"
}
```

Response Content Type

Parameters

Parameter	Value	Description	Parameter Type	Data Type
paper_id	<input type="text" value="(required)"/>	Publisher assigned identifier of a paper	path	string
publisher	<input type="text" value="aminer"/>	Publisher to perform this operation on	path	string



Swagger UI

Curl

```
curl -X GET --header "Accept: application/json" "http://babel-us-east-1.eigenfactor.org/recommendation/aminer/12345?limit=5"
```

Request URL

```
http://babel-us-east-1.eigenfactor.org/recommendation/aminer/12345?limit=5
```

Response Body

```
{
  "transaction_id": "d403eb04898d9db782f30b14c983bec6",
  "results": [
    {
      "paper_id": "235114",
      "publisher": "aminer"
    },
    {
      "paper_id": "24108",
      "publisher": "aminer"
    },
    {
      "paper_id": "121114",
      "publisher": "aminer"
    },
    {
      "paper_id": "1058620",
      "publisher": "aminer"
    },
    {

```



Frontend

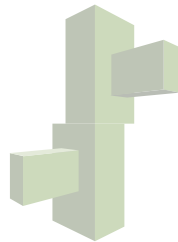
Application



Package



Deploy



AWS Elastic Bean Stalk





```
20 lines (12 sloc)  462 Bytes  Raw Blame History   
1 FROM python:3-onbuild
2
3 WORKDIR /var/app
4
5 RUN pip3 install virtualenv
6 RUN virtualenv /var/app
7
8 RUN useradd uwsgi -s /bin/false
9 RUN mkdir /var/log/uwsgi
10 RUN chown -R uwsgi:uwsgi /var/log/uwsgi
11
12 ADD . /var/app
13 RUN if [ -f /var/app/requirements.txt ]; then /var/app/bin/pip install -r /var/app/requirements.txt; fi
14
15 ENV BABEL_STAGE beta
16
17 EXPOSE 8080
18
19 CMD ["python", "./src/application.py"]
```



Frontend

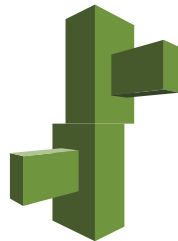
Application



Package

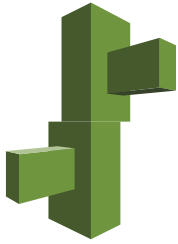


Deploy



AWS Elastic Bean Stalk





AWS Elastic Bean Stalk

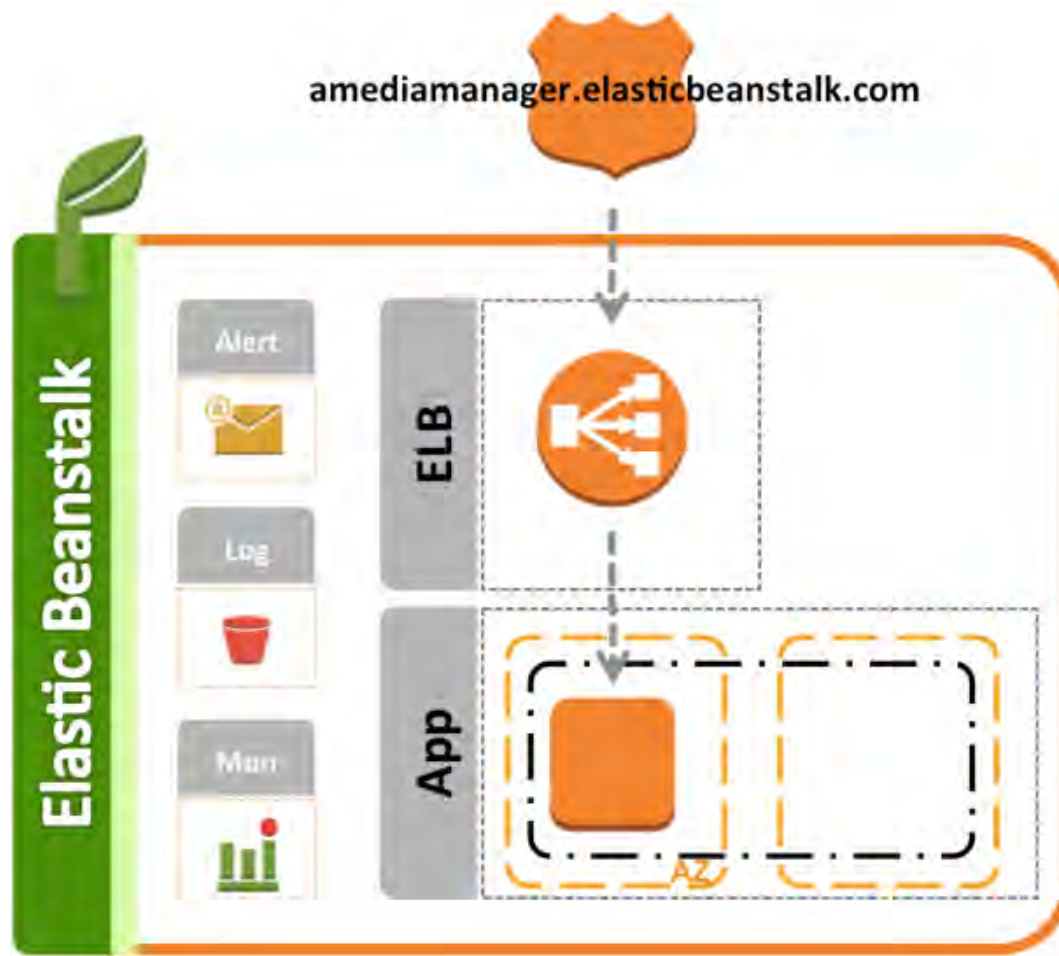
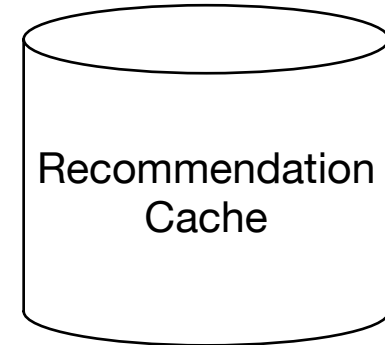


Image: [Part 1: Develop, Deploy, and Manage for Scale with Elastic Beanstalk and CloudFormation Series](#) by Evan Brown, AWS

DynamoDB

- AWS NoSQL
 - Key-value store
- Very fast (<10ms)
- Very scalable
 - Specify throughput
- Not too expensive



Issues

- Not all AWS services are created equal
 - Data Pipeline
 - Cloud Search
- Documentation
- SDK/Tooling
- Python & GIL
- Access Keys



Future Directions

- Finish backend
- Expand clients (publishers, tool developers)
- Actually get more recommenders
- Babel 3.0 – simple middleware
 - Automatically logs & add transaction info to outgoing requests



<http://babel.eigenfactor.org>

iwsmith@uw.edu

