



## **EPL646 – Advanced Topics in Databases**

# **Lecture 12**

## **Big Data Management II (NoSQL Databases / CouchDB)**

**Chapter 20: Abiteboul et. Al.  
+ <http://guide.couchdb.org/>**

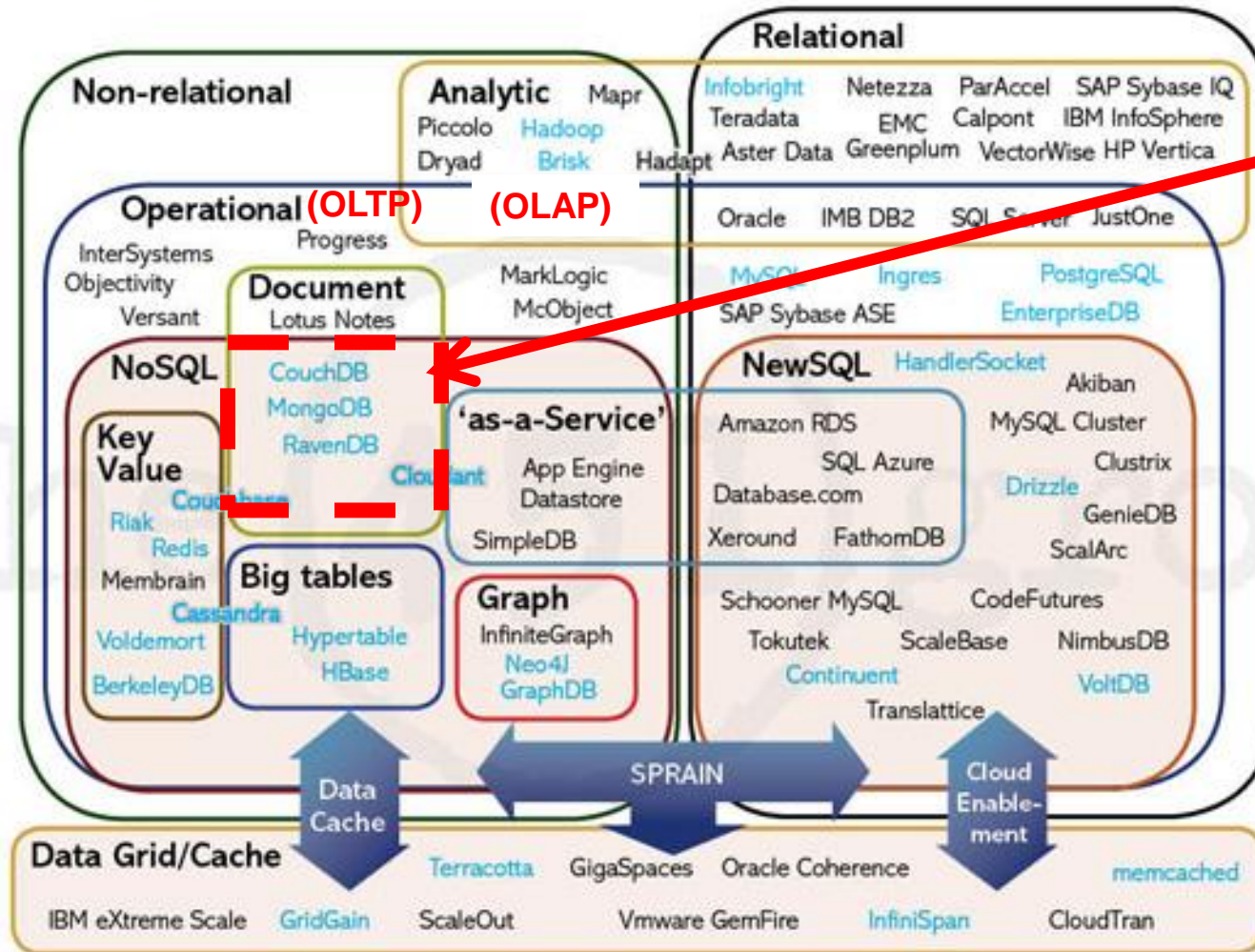
**Demetris Zeinalipour**

<http://www.cs.ucy.ac.cy/~dzeina/courses/epl646>

# EPL646: Part B



## Distributed/Web/Cloud DBs/Dstores



Lecture Focus

Venn Diagram by 451 group

<http://xeround.com/blog/2011/04/newsq-cloud-database-as-a-service>

# Lecture Outline



## (Introduction to Semi-structured Data)

- Intro to Web2.0 & JSON Data Interchange Format
- JSON Key-Value Data Model
- CouchDB: A JSON Database (written in Erlang)
  - Using Command Line CURL/ Web-based FUTON
  - CouchDB Architecture (Btrees, Filesystem, Replication)
  - REST Principles
  - Creating DBs, Adding Docs, Updating Docs, Deleting Docs, `_ID` and `_REV` issues, Multi-Version CC (MVCC)
  - Querying Data with (Materialized) Views (Map-Reduce style in Javascript)
  - Replication and Scalability Issues

# Web 2.0: The Structured Web



DBLP: <http://www.informatik.uni-trier.de/>

[ Numerous sites already allow downloading remote repositories in structured form (e.g., XML) ]

## Links

- **Computer Science Organizations:** [ACM](#) ( [DL](#) / [SIGMOD](#) / [SIGIR](#) ), [IEEE Computer Society](#) ( [DL](#) ), [IEEE Xplore](#), [IFIP](#), ...
- **Related Services:** [Google Scholar](#), [MS Academic Search](#), [CiteSeer/ CiteSeerX](#), [CS BibTeX](#) ( [DBLP](#) ), [io-port.net](#), [CoRR](#), [HAL](#), [NZ-DL](#), [Zentralblatt MATH](#), [MathSciNet](#), [Erdős Number Proj.](#), [Math Genealogy Proj.](#), [BibSonomy](#), [CiteULike](#), [ScientificCommons](#), [Libra](#), [Arnetminer](#), [RePEc](#), ...

## Schloss Dagstuhl and DBLP join forces

A joint cooperation between [Schloss Dagstuhl](#) and Trier University/DBLP aims at strengthening the documentation of research publications in Informatics in a comprehensive, transparent, and open accessible way. ...  
([news](#)) - ([project page](#))

## XML

You may download **DBLP XML records** from <http://dblp.uni-trier.de/xml/> - a [simple DTD](#) is available. The paper "[DBLP - Some Lessons Learned](#)" documents technical details of this XML file. In the appendix "[DBLP XML Requests](#)" you may find the description of a primitive DBLP API.

## New Design

On the host [Trier II](#) we are testing a new design for the DBLP website ...

**2 100 000**

DBLP now lists more than 2.1 million (> 2<sup>21</sup>) publications. More [Statistics about DBLP](#)

# JSON: Web 2.0 Data Interchange Format

## (JSON: The Fat-free XML)



- The initial vision for XML was to provide a data-**interchange language** to enable **machine-to-machine** communication.
  - However, XML is not well suited to data-interchange as the elements are **taking up to much space**.
- **JSON (JavaScript Object Notation)**
  - RFC4627: a lightweight, text-based, language-independent data interchange format.
- Web services providers nowadays offer their web services in JSON (e.g., Google APIs, Twitter API)
  - The objective of this lecture is to see how to store/query such data with a specialized document store, titled CouchDB (other: MongoDB (open), RavenDB (open))<sup>12-5</sup>

# JSON: Web 2.0 Data Interchange Format (Google Books)



## Web1.0: The Unstructured Web <http://books.google.com/>

The screenshot shows a Google search interface. The search bar contains the word "databases". Below the search bar, the results are displayed in a list format. On the left side, there are navigation options: "Web", "Images", "Maps", "Videos", "News", "Shopping", "Books", and "More". The "Books" option is highlighted in red. Below the navigation options, there are three search results for books:

- Databases: Organizing Information**  
books.google.com/books?isbn=143589426X  
Greg Roza - 2010 - Preview - More editions  
Describes how databases work and how to use tables, files, and relational databases.
- Advanced Database Systems: 10th British National Conference on ...**  
books.google.com/books?isbn=3540556931  
Peter M.D. Gray, Rob J. Lucas - 1992 - Preview - More editions  
The theme of this book is the potential of new advanced database systems. The volume presents the proceedings of the 10th British National Conference on Databases, held in Aberdeen, Scotland, in July 1992.
- Databases Illuminated**  
books.google.com/books?isbn=1449606008  
Catherine Ricardo - 2011 - Preview - More editions  
This Second Edition has been revised and updated to incorporate information about the new releases of Access 2010, Oracle 11g, and InterSystems Cache.

At the bottom of the results, there is a link for **Fuzzy Databases: Modeling, Design And Implementation** with the URL [books.google.com/books?isbn=1501402243](http://books.google.com/books?isbn=1501402243).

(content in HTML only  
apprehensible to User)

# JSON: Web 2.0 Data Interchange Format (Google Books API)



## Web2.0: The Semi-structured Web!

<https://www.googleapis.com/books/v1/volumes?q=databases>

**content in XML/JSON  
apprehensible to Computer  
(data / format decoupled)**

```
{
  "kind": "books#volumes",
  "totalItems": 899,
  "items": [
    {
      "kind": "books#volume",
      "id": "4Z6tfpuBmmgC",
      "etag": "urgGiT9QlG4",
      "selfLink": "https://www.googleapis.com/books/v1/volumes/4Z6tfpuBmmgC",
      "volumeInfo": {
        "title": "Databases",
        "subtitle": "Organizing Information",
        "authors": [
          "Greg Roza"
        ],
        "publisher": "Rosen Central",
        "publishedDate": "2010-08-15",
        "description": "Describes how databases work and how to use tables, files, and relational databases.",
        "industryIdentifiers": [
          {
            "type": "ISBN_10",
            "identifier": "143589426X"
          },
          {
            "type": "ISBN_13",
            "identifier": "9781435894266"
          }
        ],
        "pageCount": 48,
        "printType": "BOOK",
        "contentVersion": "preview-1.0.0",
        "imageLinks": {
          "smallThumbnail": "http://bks3.books.google.com/books?id=4Z6tfpuBmmgC&printsec=frontcover&img=1&zoom=5&edge=curl&img alt=Small thumbnail image",
          "thumbnail": "http://bks3.books.google.com/books?id=4Z6tfpuBmmgC&printsec=frontcover&img=1&zoom=1&edge=curl&img alt=Thumbnail image"
        },
        "language": "en",
        "previewLink": "http://books.google.com/books?id=4Z6tfpuBmmgC&printsec=frontcover&dq=databases&hl=&cd=1&source=books_api",
        "infoLink": "http://books.google.com/books?id=4Z6tfpuBmmgC&dq=databases&hl=&source=books_api"
      }
    }
  ]
}
```

*<https://www.googleapis.com/books/v1/volumes?q=flowers+inauthor:keyes&key=yourAPIKey> =>  
Provides additional details (e.g., purchase status)*

# JSON: Web 2.0 Data Interchange Format (Twitter API)



<https://twitter.com/users/dmslucy.json>

```
- {
  "id":742558014,
  "follow_request_sent":null,
  "following":null,
  "screen_name":"DMSLUCY",
  "url":"http:\\\\dmsl.cs.ucy.ac.cy\\/",
  "profile_use_background_image":true,
  "created_at":"Tue Aug 07 09:36:30 +0000 2012",
  "profile_text_color":"333333",
  "utc_offset":7200,
  "statuses_count":10,
  "default_profile_image":false,
  "verified":false,
  "name":"DMS Laboratory, UCY",
  "favourites_count":10,
  "profile_sidebar_border_color":"C0DEED",
  "friends_count":0,
  "profile_image_url_https":"https:\\\\si0.twimg.com\\profil
e_images\\2728729106\\130bc7921970a06228d1ad0d352260de_nor
mal.png",
  "description":"DMSL belongs to the Computer Science
Department at the University of Cyprus. We focus on Data
Engineering Systems and Knowledge Discovery Solutions. ",
  "profile_image_url":"http:\\\\a0.twimg.com\\profile_images
\\2728729106\\130bc7921970a06228d1ad0d352260de_normal.png"
,
```



# JSON: Web 2.0 Data Interchange Format (Google Geolocation API)



```
curl -d @request.json -H "Content-Type: application/json" -i  
"https://www.googleapis.com/geolocation/v1/geolocate?key=YOURKEY"
```

## Request Format (request.json)

```
{  
  "homeMobileCountryCode": 310,  
  "homeMobileNetworkCode": 260,  
  "radioType": "gsm",  
  "carrier": "T-Mobile",  
  "cellTowers": [  
    {  
      "cellId": 39627456,  
      "locationAreaCode": 40495,  
      "mobileCountryCode": 310,  
      "mobileNetworkCode": 260,  
      "age": 0,  
      "signalStrength": -95  
    },  
    {  
      "macAddress": "01:23:45:67:89:AB",  
      "signalStrength": 8,  
      "age": 0,  
      "signalToNoiseRatio": -65,  
      "channel": 8  
    },  
    {  
      "macAddress": "01:23:45:67:89:AC",  
      "signalStrength": 4,  
      "age": 0  
    }  
  ],  
  "wifiAccessPoints": [  
    {  
      "macAddress": "01:23:45:67:89:AB",  
      "signalStrength": 8,  
      "age": 0,  
      "signalToNoiseRatio": -65,  
      "channel": 8  
    },  
    {  
      "macAddress": "01:23:45:67:89:AC",  
      "signalStrength": 4,  
      "age": 0  
    }  
  ]  
}
```

## Response Format

The response format is also JSON.

```
{  
  "location": {  
    "latitude": 51.0,  
    "longitude": -0.1,  
  },  
  "accuracy": 1200.4,  
}
```

# JSON: Web 2.0 Data Interchange Format (Other Google APIs)



**In fact, Web2.0 Services are omnipresent!**

(Google, Twitter, Facebook, Youtube, Linkedin, ...)

<http://www.programmableweb.com/> - 7800 APIs!!! + 6800 Mashups!

The screenshot shows the Google APIs console interface. At the top, the URL is `https://code.google.com/apis`. Below the header, there are tabs for 'All (51)', 'Active (2)', 'Inactive (48)', and 'Google Cloud Platform'. The 'All services' section is active, displaying a table of services. The table has three columns: 'Service', 'Status', and 'Notes'. The services listed include Ad Exchange Buyer API, AdSense Host API, AdSense Management API, Analytics API, Audit API, BigQuery API, Blogger API v3, Books API, Calendar API, and Custom Search API. Each service has a status toggle and a note about its courtesy limit.

Service	Status	Notes
Ad Exchange Buyer API	OFF	Courtesy limit: 1,000 requests/day
AdSense Host API	<a href="#">Request access...</a>	Courtesy limit: 100,000 requests/day
AdSense Management API	OFF	Courtesy limit: 10,000 requests/day
Analytics API	OFF	Courtesy limit: 50,000 requests/day
Audit API	OFF	Courtesy limit: 10,000 requests/day
BigQuery API	OFF	Courtesy limit: 10,000 requests/day • <a href="#">Pricing</a>
Blogger API v3	<a href="#">Request access...</a>	Courtesy limit: 10,000 requests/day
Books API	ON	Courtesy limit: 1,000 requests/day
Calendar API	OFF	Courtesy limit: 10,000 requests/day
Custom Search API	OFF	Courtesy limit: 100 requests/day • <a href="#">Pricing</a>

# The JSON Key-Value Data Model



## At the core: key-value construct

Basic example:

```
"title": "The Social network"
```

Atomic data types: character strings, integers, floating-point number and Booleans (`true` or `false`). Non-string values need not be surrounded by `""`.

```
"year": 2010
```

# The JSON Key-Value Data Model



## Complex values: objects

An *object* is an unordered set of name/value pairs.

*Json does not care about types (everything is essentially text)*

The types can be distinct, and a key can only appear once.

```
{"last_name": "Fincher", "first_name": "David"}
```

A object can be used as the (complex) value component of a key-value construct:

```
"director": {  
  "last_name": "Fincher",  
  "first_name": "David",  
  "birth_date": 1962  
}
```

# The JSON Key-Value Data Model



## Complex values: arrays

An array is an ordered collection of values that need not be of the same type.

```
"actors": ["Eisenberg", "Mara", "Garfield", "Timberlake"]
```

A *document* is an object. It can be represented with an unbounded nesting of array and object constructs

```
{
  "title": "The Social network",
  "year": "2010",
  "director": {"last_name": "Fincher",
               "first_name": "David"},
  "actors": [
    {"first_name": "Jesse", "last_name": "Eisenberg"},
    {"first_name": "Rooney", "last_name": "Mara"}
  ]
}
```

# CouchDB: A JSON Database



## What is CouchDB?

*"a database that completely embraces the web"*

A system representative of the “NoSQL” trend.

- 1 a semi-structured data model, based on JSON;
  - 2 no schema;
  - 3 *structured materialized views* produced from document collections;
  - 4 views defined with the MAPREDUCE paradigm, allowing both a parallel computation and incremental maintenance of their content;
  - 5 distributed data management techniques: consistent hashing, support for data replication and reconciliation, horizontal scalability, parallel computing, etc.
- conflict resolution



## CouchDB

relax

<http://couchdb.apache.org/>

# CouchDB: FUTON Web Admin GUI



## Futon: A Web-based front-end for administering CouchDB

The screenshot shows the Futon web interface for CouchDB. The browser address bar displays `127.0.0.1:5984/_utils/`. The main content area is titled "Overview" and features a "+ Create Database ..." button. Below this is a table listing databases:

Name	Size	Number of Documents	Update Seq
<a href="#">_replicator</a>	4.1 KB	1	1
<a href="#">_users</a>	4.1 KB	1	1
<a href="#">books</a>	3.7 MB	1557	1563
<a href="#">movies</a>	4.1 KB	1	1
<a href="#">twitter</a>	4.1 KB	1	1

At the bottom of the table, it says "Showing 1-5 of 5 databases" and includes navigation links: "← Previous Page", "Rows per page: 10", and "Next Page →".

The right sidebar contains the CouchDB logo with the tagline "relax" and a "Tools" menu with the following items:

- Overview
- Configuration
- Replicator
- Status
- Verify Installation

Below the menu, there is a section for developers: "For Developers: Test Suite".

# CouchDB: FUTON Web Admin GUI



## Editing records (documents) with Futon

← → ↻ ⬆ 127.0.0.1:5984/\_utils/database.html?books/\_all\_docs

Overview > books

+ New Document ⓘ Security... Jump to:  View: All documents ⌵ Stale views

⊗ Compact & Cleanup...  
⊗ Delete Database...

Key ▲	Value
"_design/examples" ID: _design/examples	{rev: "4-ab4f0d8f5340146bdcb32a78e137f0a9"}
"book1.json" ID: book1.json	{rev: "1-410c67caca526b476abc72e73b003605"}
"book10.json" ID: book10.json	{rev: "1-d0cc2ae0ab3211314a65a5c5244df221"}
"book100.json" ID: book100.json	{rev: "1-2cfe83eea8cad920cfd66755ac78b46f"}
"book1000.json"	{rev: "1-7681da7415571ed392393b99734c86d8"}



# CouchDB in a Nutshell



## CouchDB in a nutshell

A **document**, **web-oriented** data system.

**Document oriented.** Document are complex and autonomous pieces of information. Can store files, functions, any type of media. But no references.

Typical functionalities of document application: versioning, replication, synchronization, restructuring.

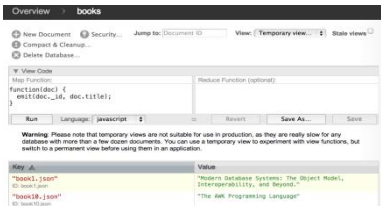
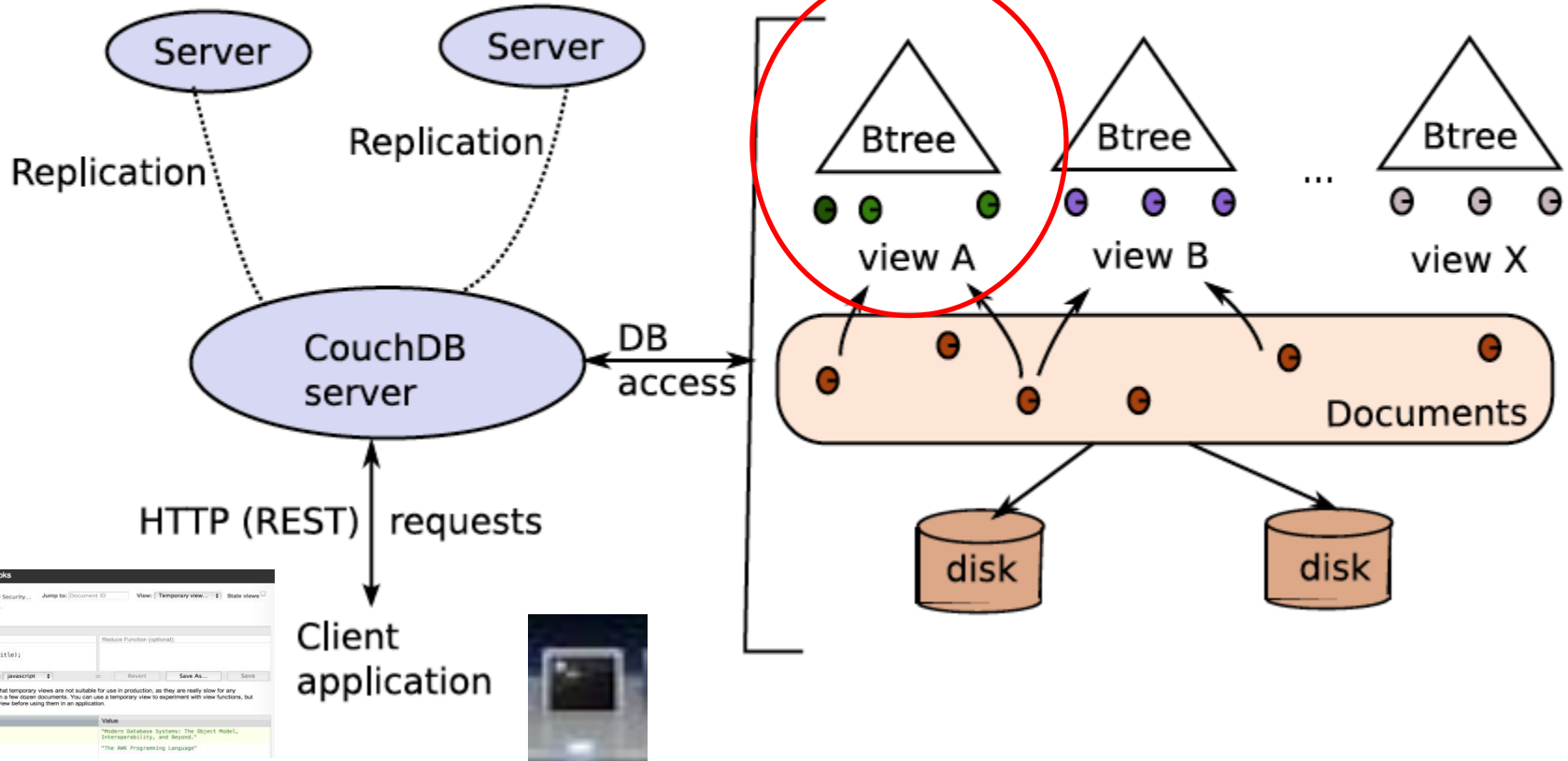
**Web-oriented.** A document is a **resource** in the Web sense – it has a URI, and can be manipulated via HTTP (REST architecture).

# CouchDB: A JSON Database (Architecture)



## Architecture of CouchDB

**B+tree Key: [key,docid]**  
 (Materialized view => on update tree is updated as well)



# CouchDB: Filesystem Layout (Datastores and Materialized Views)



## Index of /Users/dzeina/Li

Name	Size	Date
[parent directory]		
.books_design/		10/29/12 11
.delete/		10/23/12 3
.movies_design/		10/23/12 4
_replicator.couch	4.1 kB	10/23/12 9
_users.couch	4.1 kB	10/23/12 9:40:05 AM
books.couch	3.7 MB	10/29/12 11:41:48 AM
booksreplica.couch	764 kB	10/29/12 11:41:48 AM
movies.couch	4.1 kB	10/23/12 3:40:27 PM
twitter.couch	4.1 kB	10/23/12 10:50:02 AM

## Index of /Users/dzeina/Library/Application S

Name	Size	Date Modified
[parent directory]		
1fb02d640d642a6272bfa5334d2e3f42.view	188 kB	10/29/12 11:18:59 AM
20ab140f1492382baf3aafbab426f2d7.view	56.1 kB	10/23/12 4:12:14 PM
33abe94d1a00f46506f2c0b015540db1.view	8.1 kB	10/29/12 11:19:29 AM
3b5f338a51e55052f5513da8a6bd64a8.view	216 kB	10/23/12 4:52:51 PM
451385dfae51196393f2ceb3a2b780fa.view	4.1 kB	10/29/12 11:19:29 AM
45b8fde80881d168a44595ff5cc90ea1.view	16.1 kB	10/29/12 11:48:33 AM
5ccddad630dd20ffd2b6fb4833080772.view	4.1 kB	10/23/12 4:58:52 PM
64a3e2d25cf3876b0fa55d78c50d7f0b.view	72.1 kB	10/29/12 11:48:33 AM
6ba736a20feacba97c220a4b7f3e02e3.view	132 kB	10/29/12 11:45:02 AM
6da619412b8ffa95126c11d2b21d27cc.view	504 kB	10/23/12 4:09:13 PM
6fac50850cee7de2f185090669defd68.view	4.1 kB	10/23/12 11:21:22 PM
70419c9fa523e611895f53b01ac694e0.view	204 kB	10/23/12 4:59:53 PM