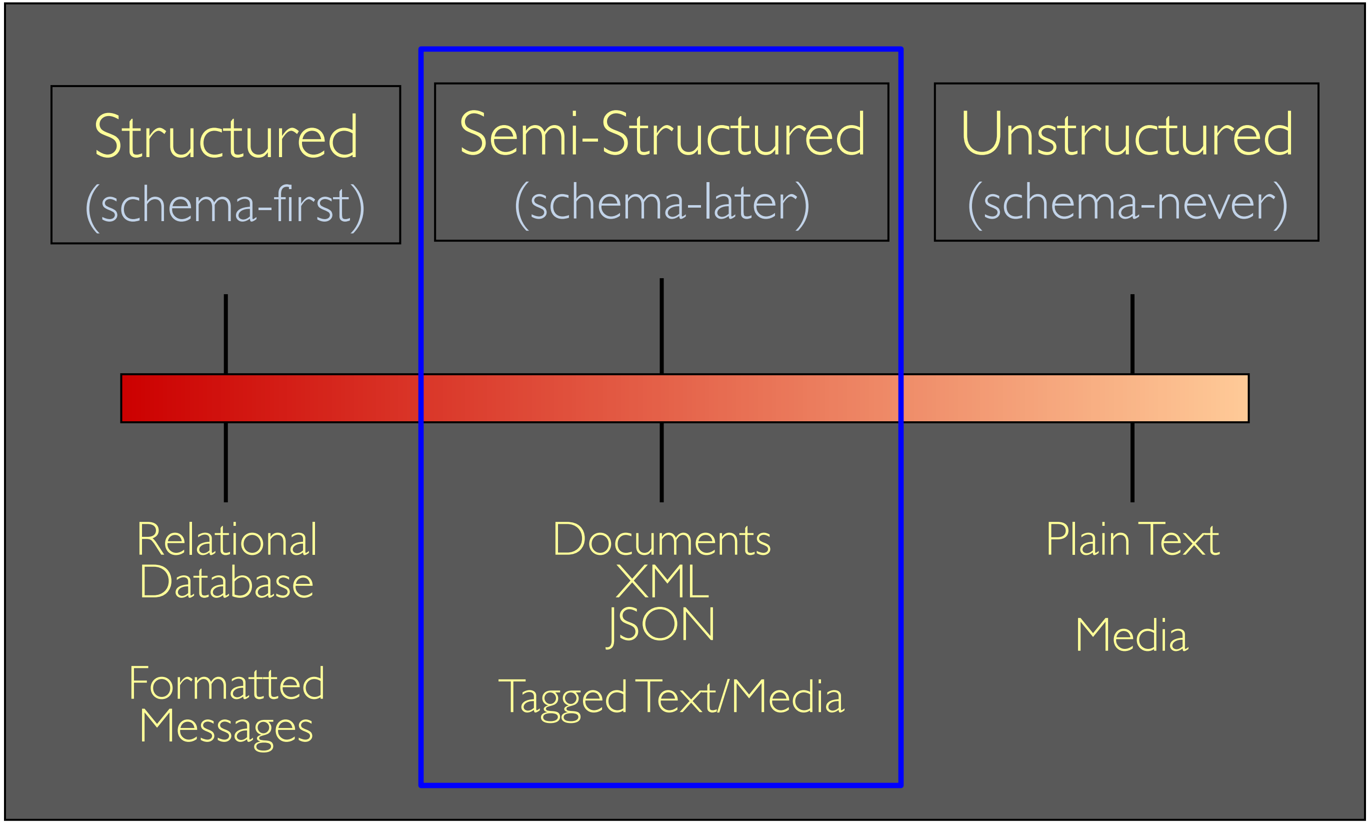
The Structure Spectrum



Structured

(schema-first)

Semi-Structured

(schema-later)

Unstructured

(schema-never)

Relational Database

Formatted Messages

Documents

XML

JSON

Tagged Text/Media

Plain Text

Media

Lab

[redacted]

Semi-Structured Log Files

- Created by `printf` statements in server processes:
  - » Web, database, network file servers, operating system components

- Human-readable text format files
  - » Very rarely actually read by a human
  - » Can store/archive in binary or compressed format

- Format published or "defined" by code
  - » Can be very difficult to parse

[redacted]

Recall: Apache Web Server Log

```
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif HTTP/
1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif
HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif HTTP/
1.0" 304 0
ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET /images/launch-
logo.gif HTTP/1.0" 200 1713
uplherc.upl.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/WORLD-logosmall.gif HTTP/
1.0" 304 0
slppp6.intermind.net - - [01/Aug/1995:00:00:10 -0400] "GET /history/skylab/skylab.htm[redacted]
HTTP/1.0" 200 1687
piweba4y.prodigy.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/launchmedium.gif
HTTP/1.0" 200 11853
tampico.usc.edu - - [14/Aug/1995:22:57:13 -0400] "GET /welcome.html HTTP/1.0" 200 79
```

Apache Web Server Log Format

- Apache Common Log Format specifies log file format

- Example line from log file:

  »
  ```
  127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0"
  200 1839
  ```

- Components:

  » `127.0.0.1` *Client IP address*

Apache Web Server Log Format

- Apache Common Log Format specifies log file format
- Example line from log file:
  - »
```
127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0"
200 1839
```
- Components:
  - » - *User identity from remote machine*

    (hyphen means not available)
  - » - *User identity from local logon*

    (hyphen means not available)

[redacted]

Apache Web Server Log Format

- Apache Common Log Format specifies log file format

- Example line from log file:

  »
  ```
  127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0"
  200 1839
  ```

- Components:

  » `[01/Aug/1995:00:00:01 -0400]`

  Request time

Apache Web Server Log Format

- Apache Common Log Format specifies log file format
- Example line from log file:
  - »
    ```
    127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0"
    200 1839
    ```
- Components:
  - » `"GET /images/launch-logo.gif HTTP/1.0"`

    Client request

    - Request method (e.g., GET, POST, etc.)
    - Endpoint (a Uniform Resource Identifier)
    - Client protocol version

[redacted]

Apache Web Server Log Format

- Apache Common Log Format specifies log file format

- Example line from log file:

  »
  ```
  127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0"
  200 1839
  ```

- Components:

  » **`200`** *Status code the server sent back to the client*

  - OK response (2xx), others: 3xx, 4xx, 5xx

  » **`1839`** *Size of the object returned to client*

  "-" if no content returned, ***<u>or sometimes 0</u>***

[redacted]

Lab: Explore Web Server Access Log

- NASA HTTP server access log
  - » http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html
- Log covers 21 days (1 Aug, 3 Aug – 22 Aug 1995)
- Log includes 1,043,177 requests
- Log partially cleaned for you
  - » Removed some very hard to parse requests

[redacted]

Some Log Analysis Questions

- Overall:
  - » What are the statistics for content being returned? Sizes, statuses?
  - » What are the types of return codes?
  - » How many 404 (page not found) errors are there?

- Temporal:
  - » How many unique hosts per day?
  - » How many requests per day?
  - » On average, how many requests per host?
  - » How many 404 errors per day?

[redacted]