



Harnessing Flexible Data in the Cloud

Matthias Brantner (brantner@28msec.com)



msec

Demo



JSON

SportsML

SportsML

MLB WatchLatest news and updates

HomeScoresSchedulesStandingsStatsMovementsPlayersTeams



Dodgers celebrate 50 years at Dodger Stadium
The Dodgers celebrate 50 years at Dodger Stadium -- inspired in part by Midnight Lights (Jimmy Deibold)

Latest News

Red Sox owner confirms team meetings, no issues with Valentine
Boston Red Sox owner John Henry said the team held meetings last month in New York to discuss the difficult season in general, but that the reported grumbling from players about manager Bobby Valentine was never an issue.

National League Game Capsules
Greg Dobbs, Austin Kearns and Bryan Petersen each knocked in two runs, as the Miami Marlins beat the Philadelphia Phillies, 9-2, to avoid a three-game sweep. Miami came into Wednesday's game with a 27-inning scoreless streak after being blanked in three consecutive games for the first time in franchise history. The streak reached 30, the longest such stretch in club history, before Dobbs hit an RBI single in

Jennings not in Wednesday's lineup
Desmond Jennings is not in Wednesday's lineup against the Mariners.

Mets option C Johnson to Triple-A
Mets optioned C Rob Johnson to Triple-A Buffalo.

Scutaro not in Wednesday's lineup
Marco Scutaro is not in Wednesday's lineup against the Nationals.

Batting AvgHome RunsRun Batted InERA

#	Player	Team	Avg
1	Mike Trout		0.34
2	Miguel Cabrera		0.324
3	Derek Jeter		0.318
4	Alex Rios		0.318
5	Joe Mauer		0.316
6	Paul Konerko		0.316
7	David Ortiz		0.316
8	Robinson Cano		0.315
9	Austin Jackson		0.314
10	Ben Revere		0.314
11	Prince Fielder		0.31
12	Adrian Gonzalez		0.309
13	Adrian Beltré		0.304
14	Billy Butler		0.302
15	A.J. Pierzynski		0.299
16	Alcides Escobar		0.298
17	Adam Jones		0.298
18	Elvis Andrus		0.297
19	Torii Hunter		0.295
20	Alex Gordon		0.294
21	Edwin Encarnación		0.293
22	Ryan Doumit		0.292
23	Michael Brantley		0.291

Latest Results

Philadelphia Phillies 4 @ Miami Marlins 0

Texas Rangers 2 @ New York Yankees 8

San Diego Padres 4 @ Atlanta Braves 1

Twitter Tracker

@SFGiants

@Odonion17 again, this is for travel costs for the families of the players (flights, hotel, dinner, etc)

@SFGiants

One of the most popular events at AT&T Park is the annual slumber party - Check out what you missed last week: <http://t.co/GLgnZK5>

@SFGiants

@pickwick in 2013 when it is announced, you can sign up for spots, typically in March.

@SFGiants

@BFUless no one is arguing, was just clarifying

Text

NITF
(XML)

SportsML

JSON

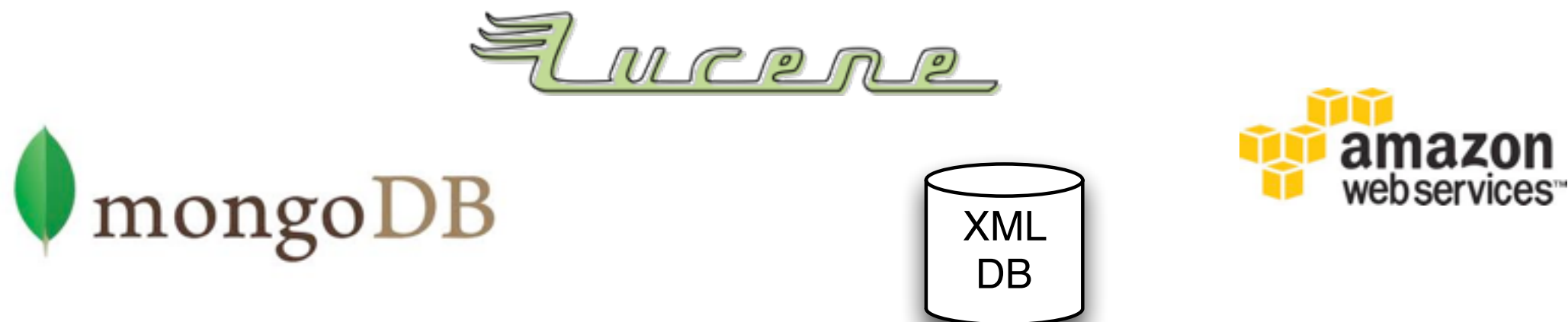
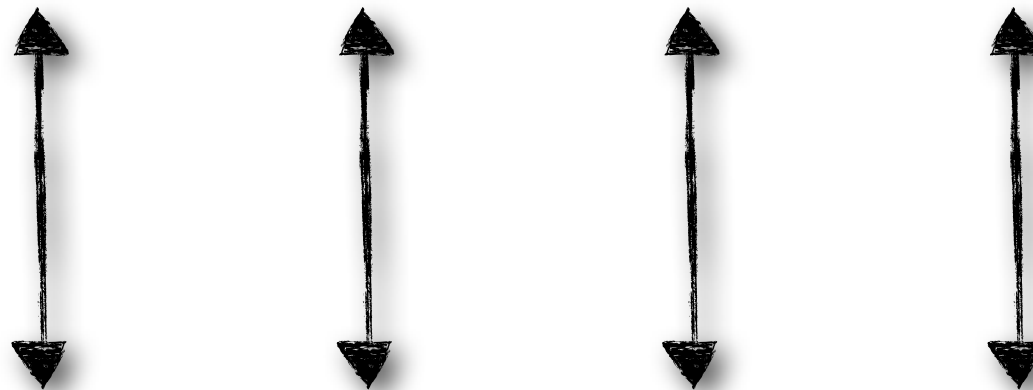
Thursday, August 23, 12

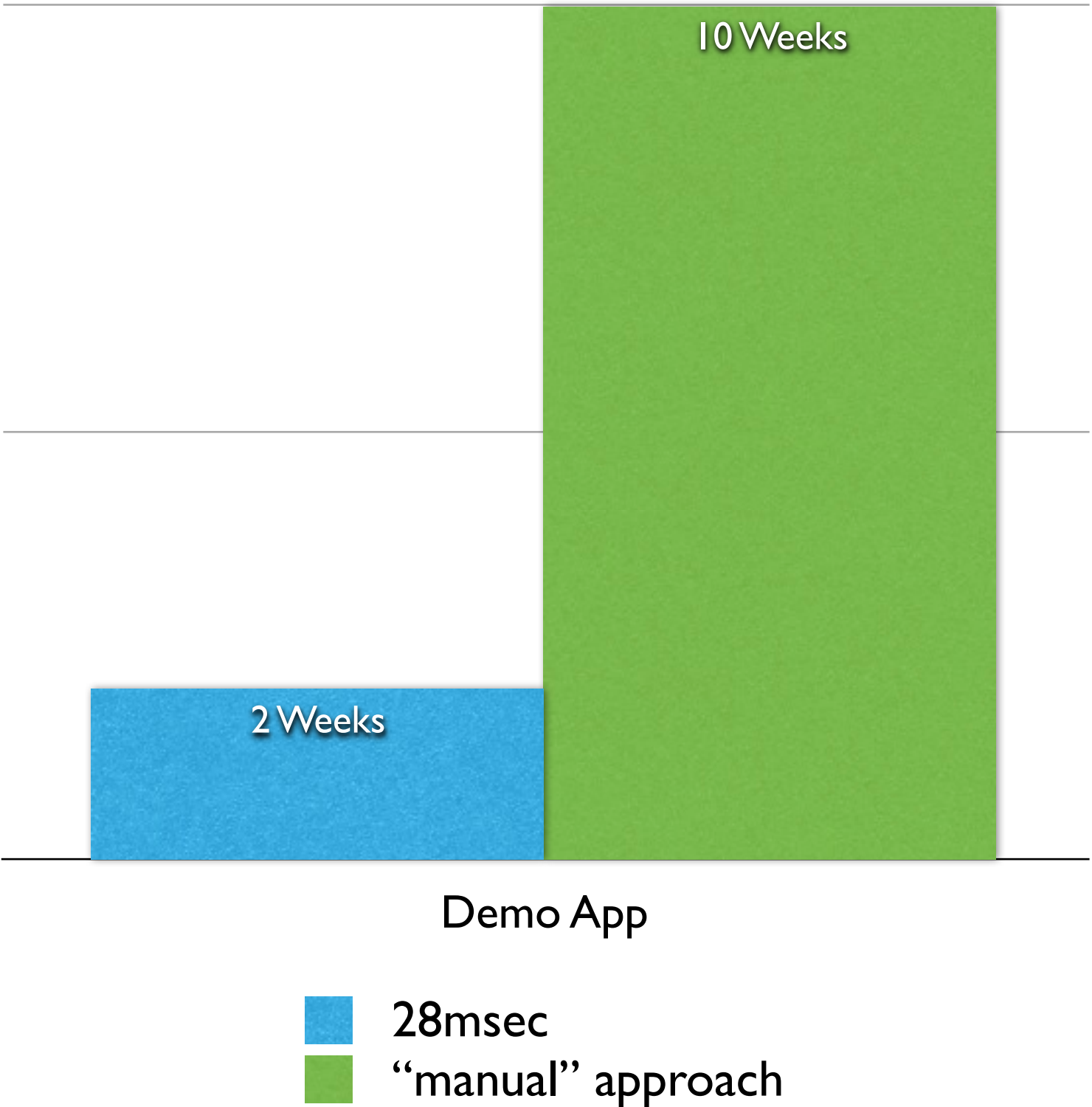
- store
- query & update
- full-text search
- complex queries
- transform

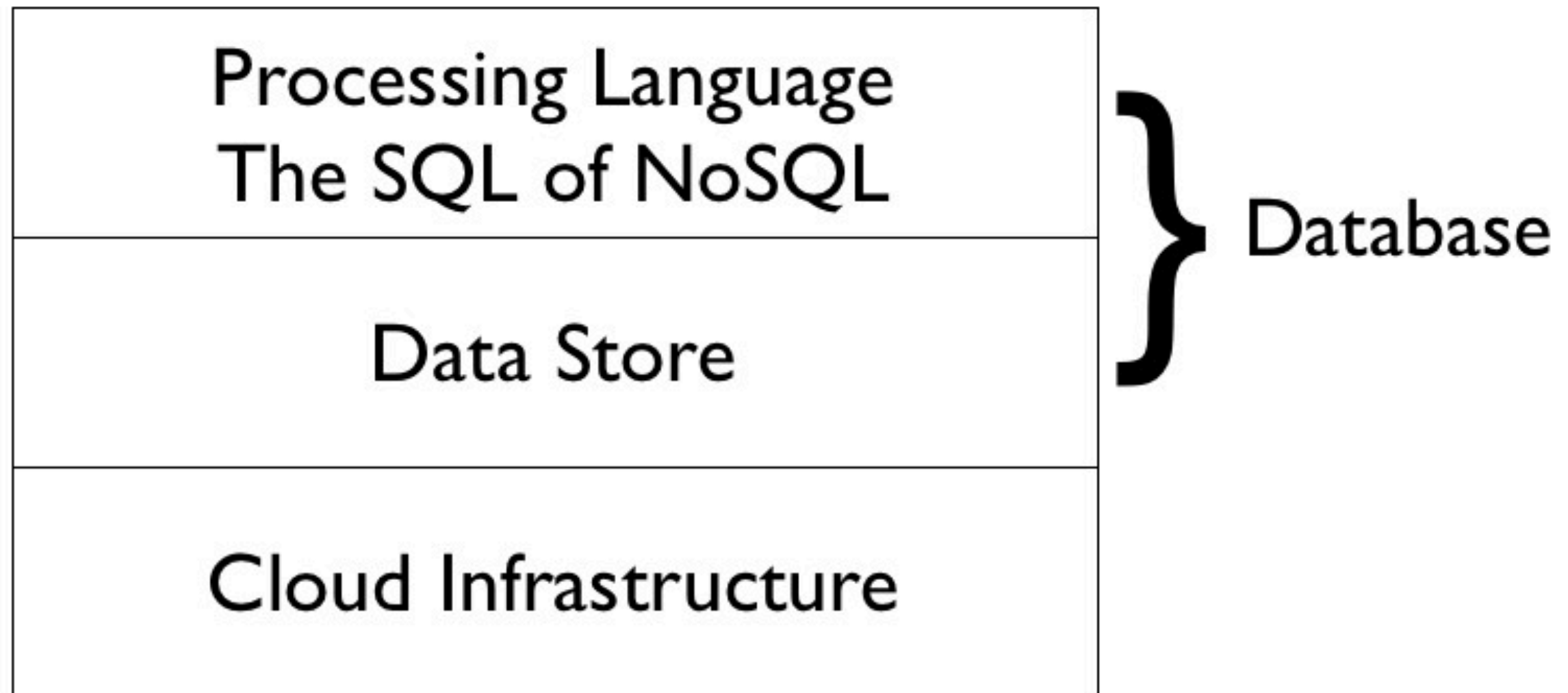
reliability, availability, performance, and scalability

Possible Solution Stack

28 msec







JSONiq					
functional	declarative	scripting	query	full-text	update
XML / JSON / Text					


```
1  for $zip in db:collection("zips")
2  group by $zip("state")
3  let $count := count($zip)
4  order by $count descending
5  return { "state" : $state, "count" : $count }
```

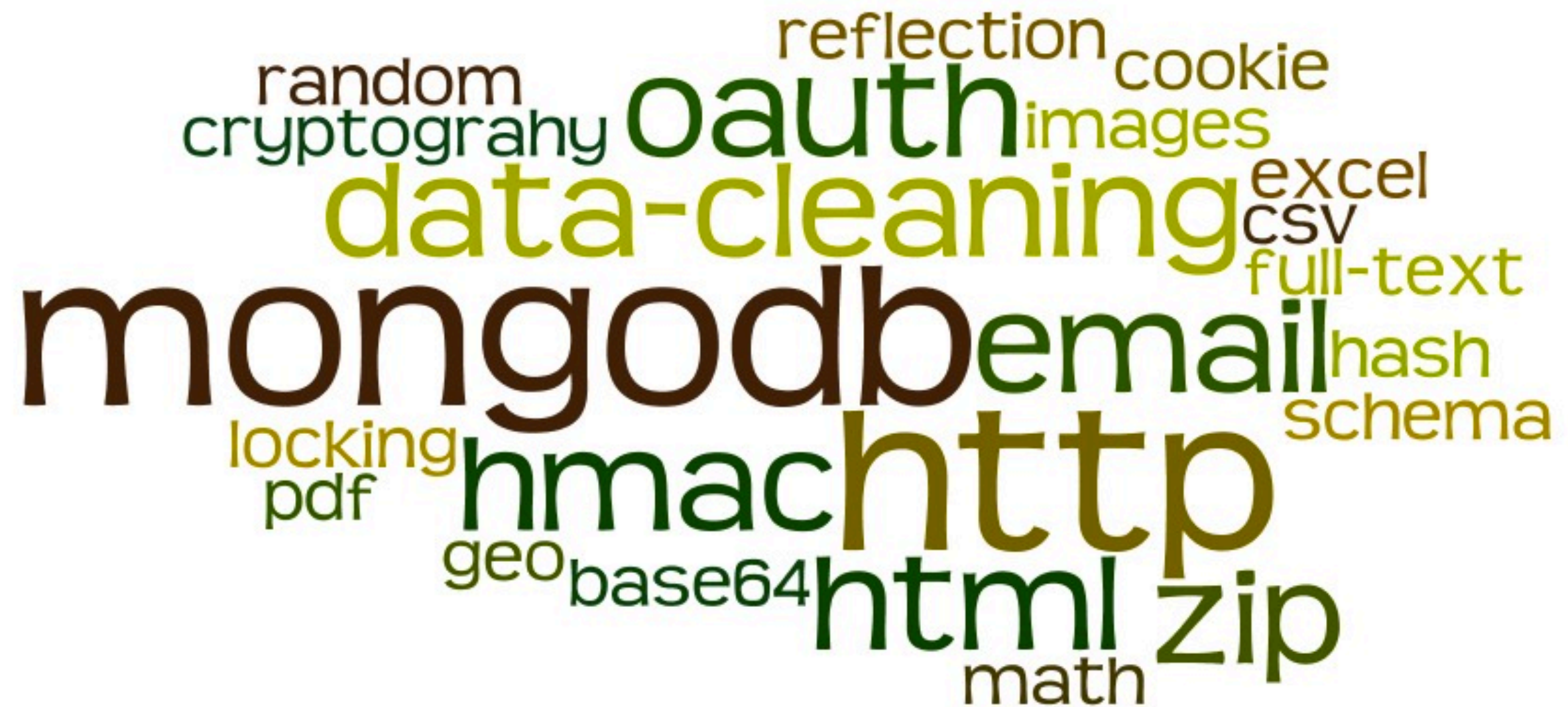
```
{ "state" : "TX", "count" : 1676 }
{ "state" : "NY", "count" : 1596 }
{ "state" : "CA", "count" : 1500 }
...
```

Example (2)

28 msec

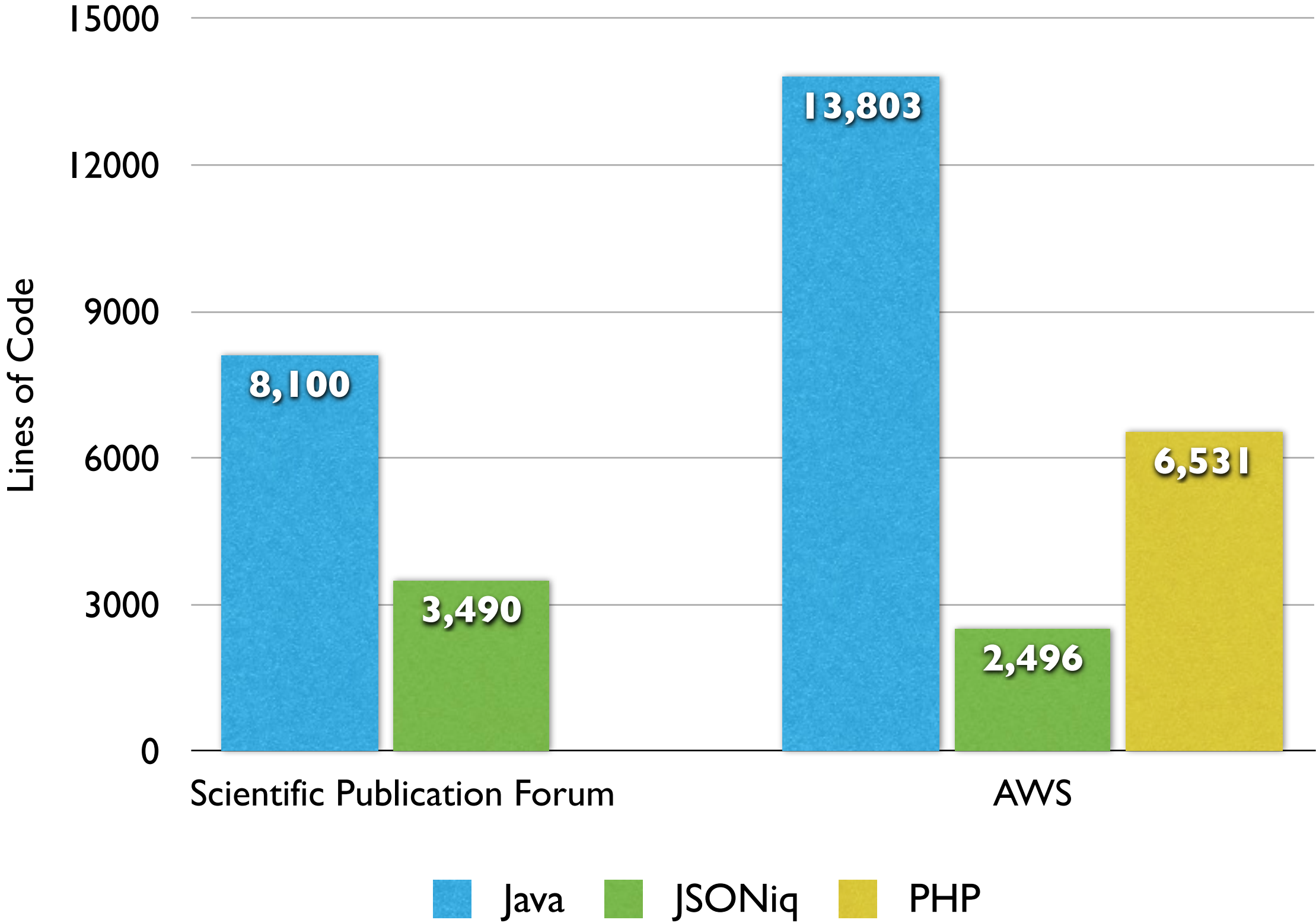
```
1 let $search-result := jn:parse-json(  
2   http:get-text("twitter.com/search.json?q=.")[2])  
3 for $result in jn:members($search-result("results"))  
4 let $text := $result("text")  
5 for $token in ft:tokenize-string($text)  
6 where not(ft:is-stop-word($token))  
7 let $lc := ft:strip-diacritics(lower-case($token))  
8 group by $lc  
9 let $count := count($token)  
10 order by $count descending  
11 return { "token" : $lc, "count" : $count }
```

```
{ "token" : "hard", "count" : 3 }  
{ "token" : "truths", "count" : 3 }  
{ "token" : "sql", "count" : 3 }  
{ "token" : "revolution", "count" : 2 }
```



A word cloud of Python modules, with 'mongodb' and 'http' being the largest. Other prominent words include 'data-cleaning', 'oauth', 'hmac', 'html', 'zip', 'reflection', 'cookie', 'images', 'excel', 'csv', 'full-text', 'hash', 'schema', 'random', 'cryptogrady', 'locking', 'pdf', 'geo', 'base64', 'math', and 'schema'.

reflection cookie
random cryptogrady oauth images
data-cleaning excel csv
mongodb email hash schema
locking pdf hmac http
geo base64 html zip
math





mongoDB

document store

performant

high availability

scalable

atomic updates

JSON / XML

JSONiq
Collection

```
{  
  "name" : "Matthias"  
  "dob" : 1978-09-12  
}
```

```
<person>  
  <name>Matthias</name>  
  <dob>1978-09-12</dob>  
</person>
```



MongoDB
Collection

```
{  
  "name" : "Matthias"  
  "dob" : 1978-09-12    ...  
}
```

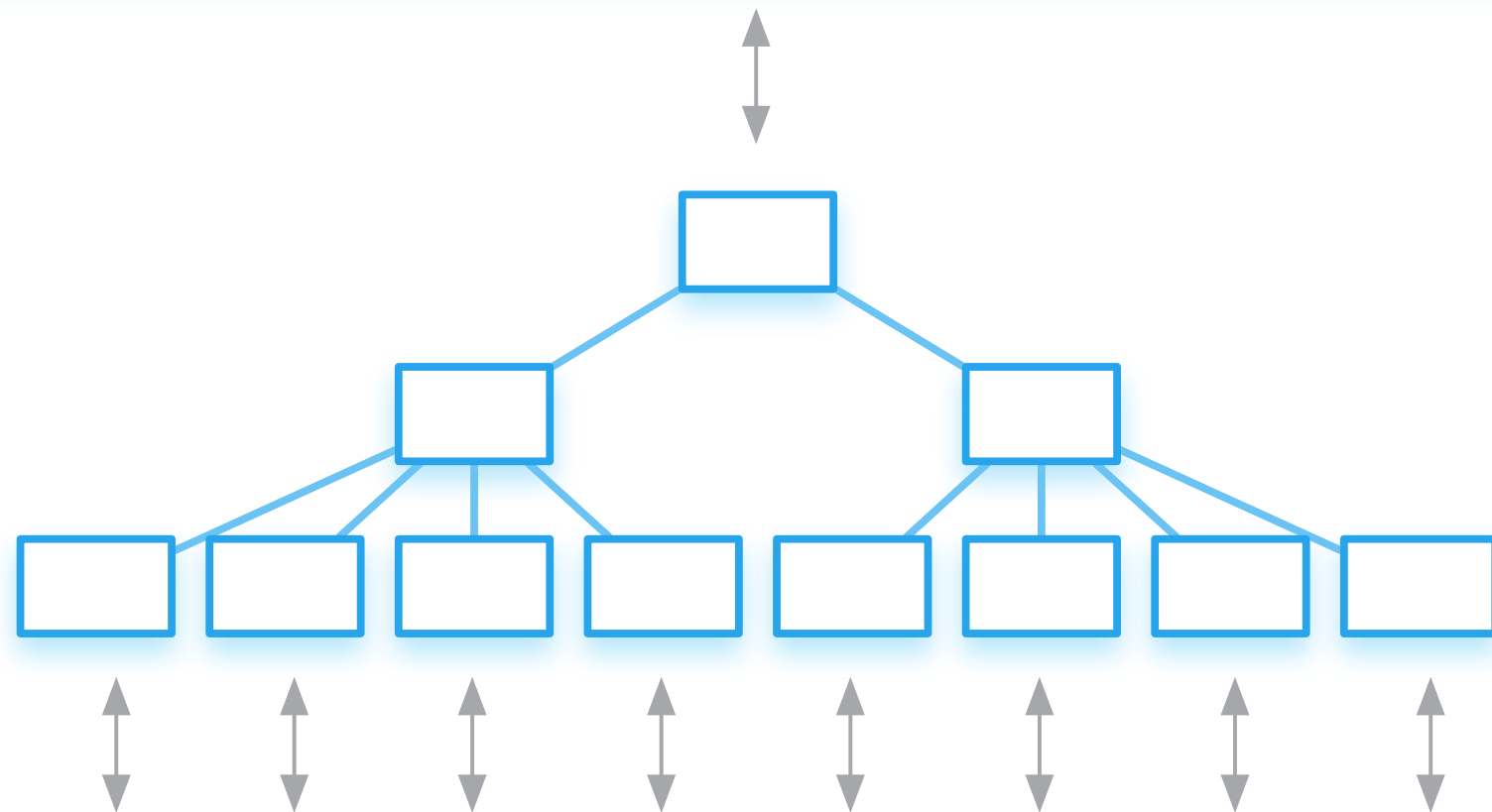
```
<person>  
  <name>Matthias</name>  
  <dob>1978-09-12</dob>  
</person>
```

JSONiq
Collection

```
{  
  "name" : "Matthias"  
  "dob" : 1978-09-12  
}
```

```
<person>  
  <name>Matthias</name>  
  <dob>1978-09-12</dob>  
</person>
```

Index (BTree)



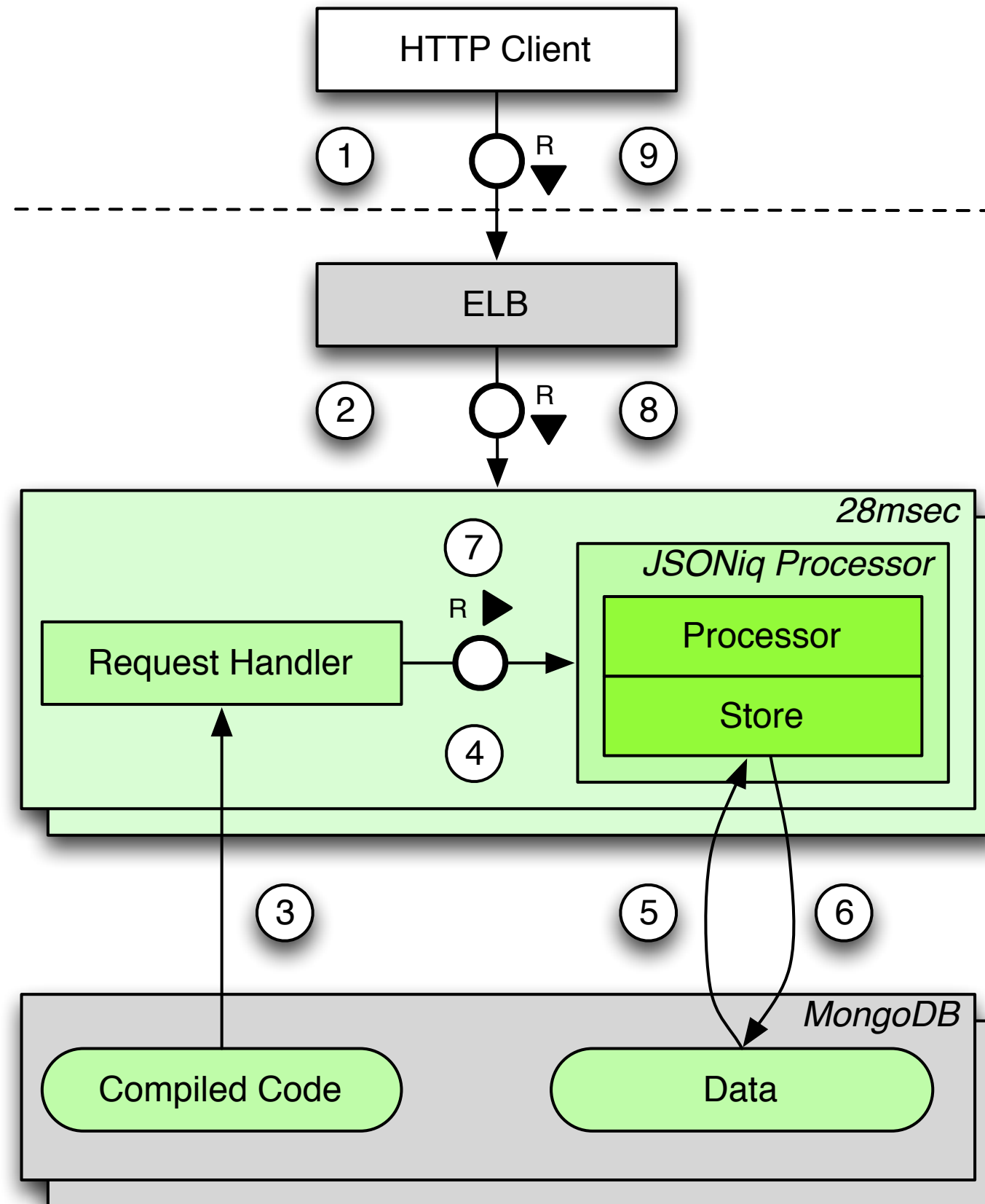
MongoDB
Collection

```
{  
  "name" : "Matthias"  
  "dob" : 1978-09-12  
}
```

```
<person>  
  <name>Matthias</name>  
  <dob>1978-09-12</dob>  
</person>
```


Cloud Infrastructure - Request Processing

28 msec



- JSONiq to process flexible data
- using MongoDB as data store and index

Database

- up to 5x more (cost) effective
 - scalable infrastructure on AWS
-

- 11:00 am
 - Jonathan Robie (EMC)
- 11:45am
 - Chris Hillery (FLWOR Foundation)



Thank You!

Questions?

Matthias Brantner (brantner@28msec.com)

Visit us at our booth!