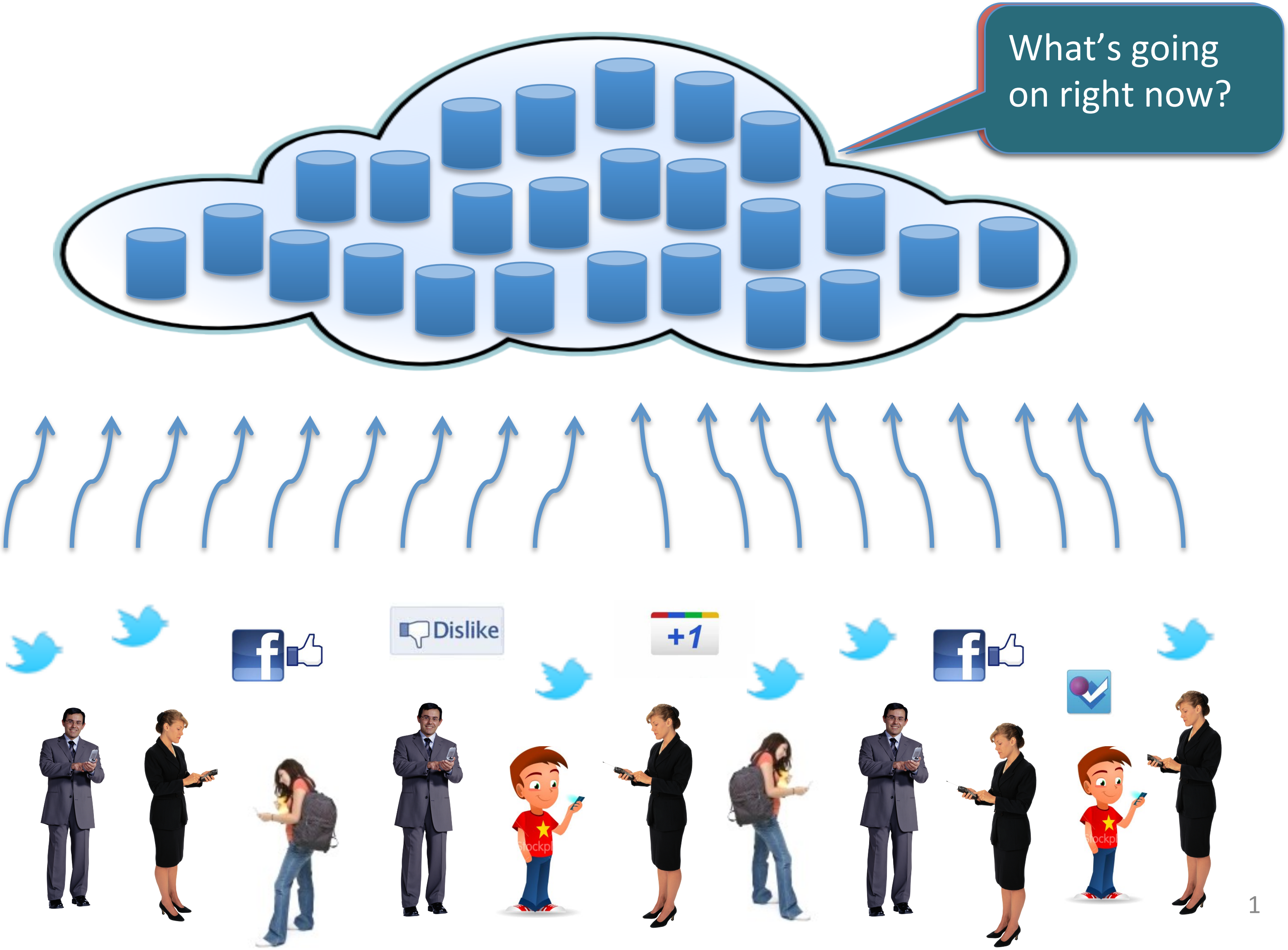
Big Data / Web Warehousing



What’s going on right now?

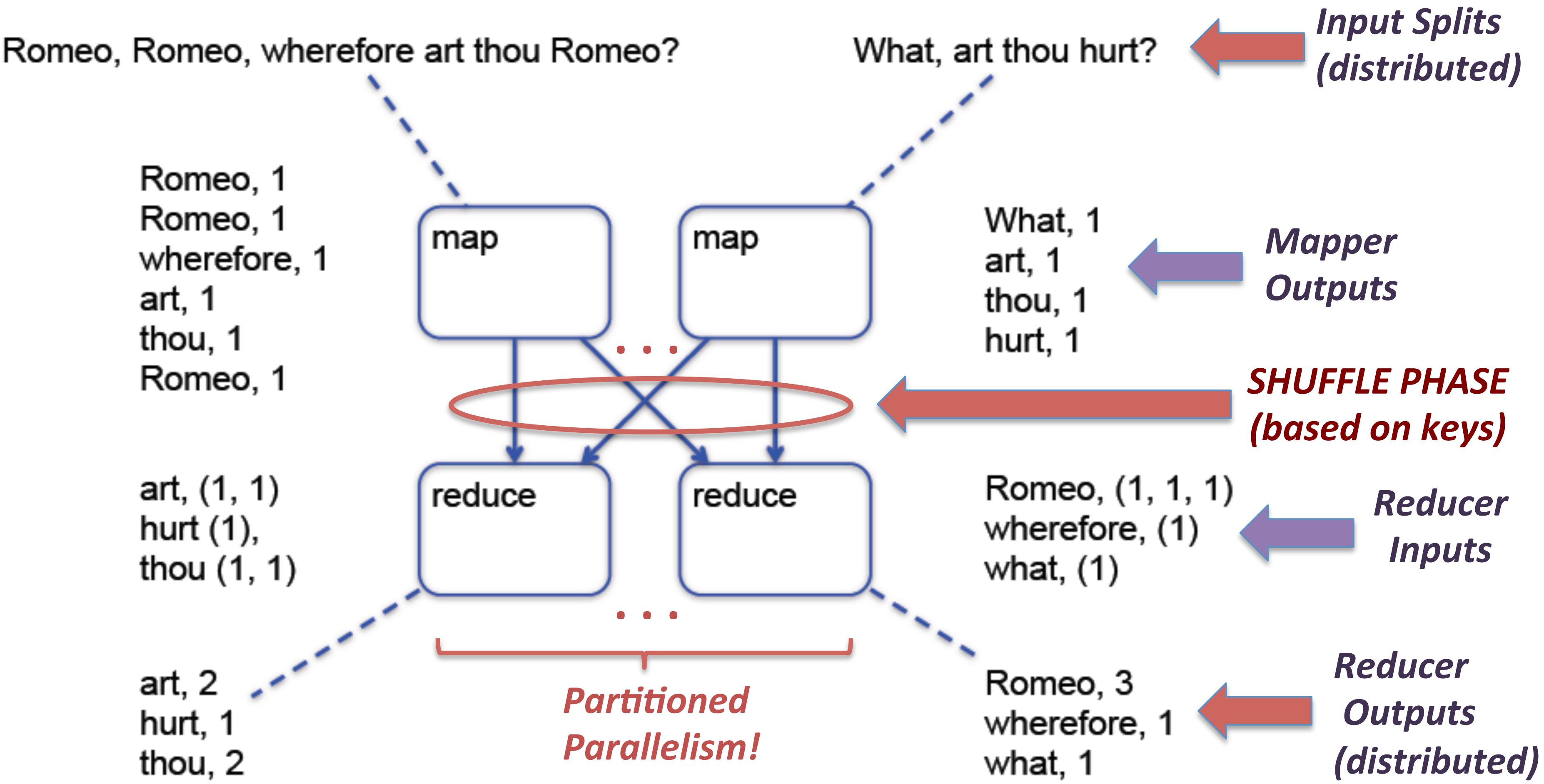
UCIRVINE

http://isg.ics.uci.edu

“Big Data” History

- Late 1990’s brought a need to index and query the rapidly exploding content of the Web
  - DB technology tried but failed (*e.g.*, Inktomi)
  - Google, Yahoo! *et al* needed to do something
- Google responded by laying a new foundation
  - Google File System (GFS)
    - OS-level byte stream files spanning 1000’s of machines
    - Three-way replication for fault-tolerance (availability)
  - MapReduce (MR) programming model
    - User functions: Map and Reduce (and optionally Combine)
    - *“Parallel programming for dummies”* – MR runtime does the heavy lifting via partitioned parallelism

(MapReduce: *Word Count* Example)



Romeo, Romeo, wherefore art thou Romeo?

What, art thou hurt?

Input Splits (distributed)

Romeo, 1
Romeo, 1
wherefore, 1
art, 1
thou, 1
Romeo, 1

What, 1
art, 1
thou, 1
hurt, 1

Mapper Outputs

SHUFFLE PHASE (based on keys)

art, (1, 1)
hurt (1),
thou (1, 1)

Romeo, (1, 1, 1)
wherefore, (1)
what, (1)

Reducer Inputs

art, 2
hurt, 1
thou, 2

Partitioned Parallelism!

Romeo, 3
wherefore, 1
what, 1

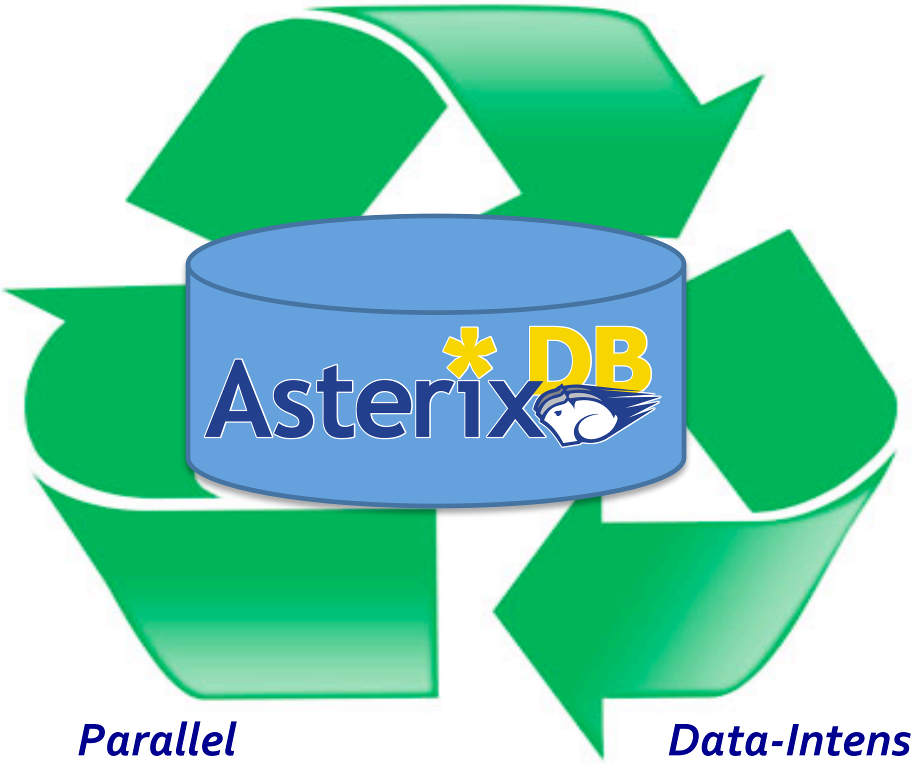
Reducer Outputs (distributed)

Today's "Big Data" Tangle

mongoDB
(Pig)
HIVE
Cassandra
ly scalable, eventually consistent, distributed, structured key-value store.
hadoop
MySQL
H·BASE

AsterixDB: “One Size Fits a Bunch”

Semistructured Data Management



Parallel Database Systems

Data-Intensive Computing

BDMS Desiderata:

- Flexible data model
- Efficient runtime
- Full query capability
- Cost proportional to task at hand (!)
- Designed for continuous data ingestion
- Support today’s “Big Data data types”

⋮

Project Goals

- Build a new Big Data Management System (BDMS)
  - Run on large commodity clusters
  - Handle mass quantities of semistructured data
  - Openly *layered*, for selective reuse by others
  - Share with the community via *open source*
- Conduct scalable information systems research, e.g.,
  - Large-scale query processing and workload management
  - Highly scalable storage and index management
  - Fuzzy matching, spatial data, date/time data (all in parallel)
  - Novel support for "fast data" (both in and out)
- Train next generation of "Big Data" graduates

ASTERIX Data Model (ADM)

```
create dataverse TinySocial;
use dataverse TinySocial;

create type MugshotUserType as {
    id: int32,
    alias: string,
    name: string,
    user-since: datetime,
    address: {
        street: string,
        city: string,
        state: string,
        zip: string,
        country: string
    },
    friend-ids: {{ int32 }},
    employment: [EmploymentType]
}
```

```
create type EmploymentType as open {
    organization-name: string,
    start-date: date,
    end-date: date?
}

create dataset MugshotUsers(MugshotUserType)
                            primary key id;
```

Highlights include:

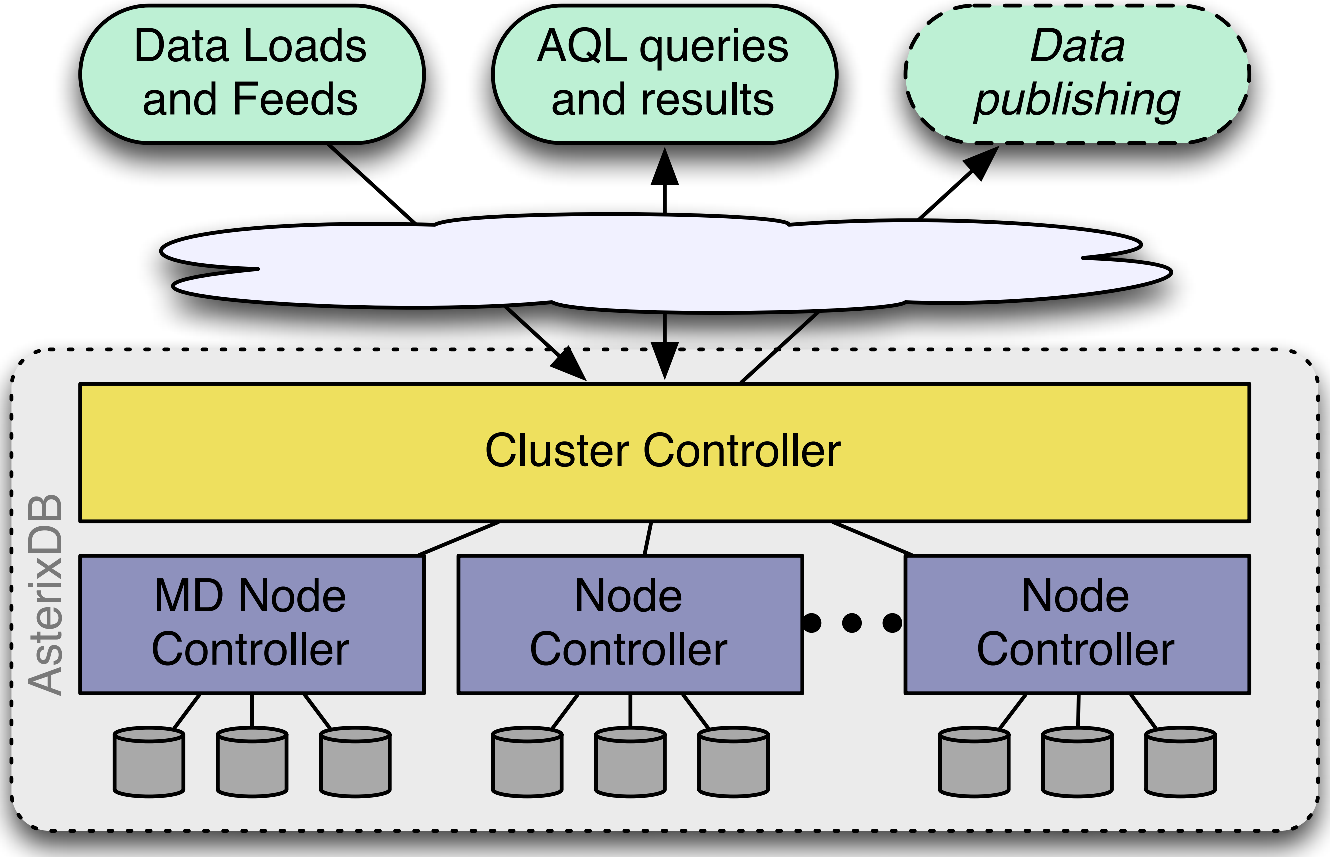
- JSON++ based data model
- Rich type support (spatial, temporal, …)
- Records, lists, bags
- ***Open vs. closed types***

ASTERIX Query Language (AQL)

- *Ex:* List the user name and messages sent by those users who joined the Mugshot social network in a certain time window:

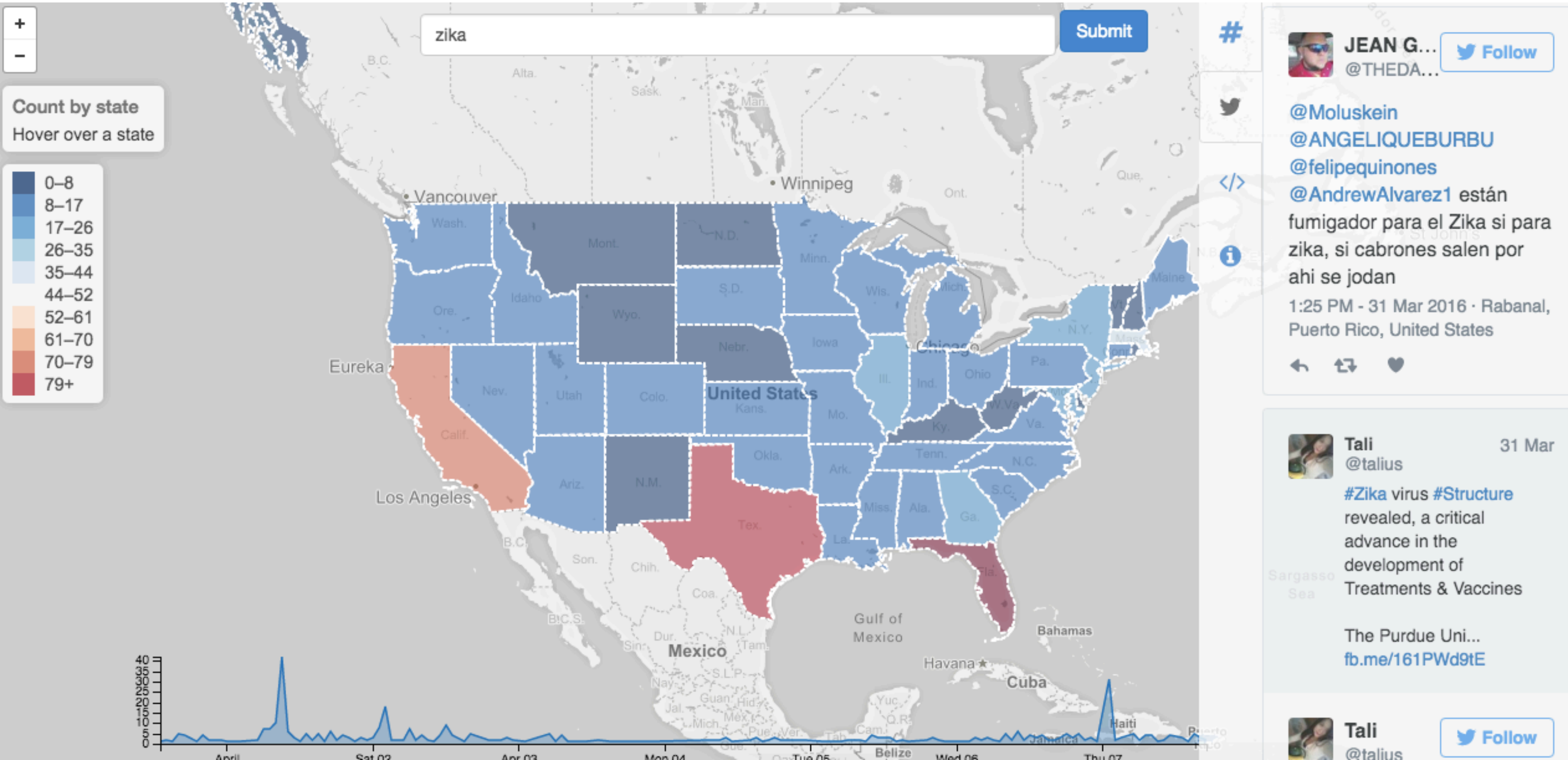
```
for $user in dataset MugshotUsers
where $user.user-since >= datetime('2010-07-22T00:00:00')
  and $user.user-since <= datetime('2012-07-29T23:59:59')
return {
  "uname" : $user.name,
  "messages" :
      for $message in dataset MugshotMessages
      where $message.author-id = $user.id
      return $message.message
};
```

AsterixDB System Overview



A prototype: interactive analytics and visualization of large data sets

http://cloudberry.ics.uci.edu/



zika

Submit

Count by state

Hover over a state

- 0–8
- 8–17
- 17–26
- 26–35
- 35–44
- 44–52
- 52–61
- 61–70
- 70–79
- 79+

JEAN G… @THEDA… Follow

@Moluskein @ANGELIQUEBURBU @felipequinones @AndrewAlvarez1 están fumigador para el Zika si para zika, si cabrones salen por ahi se jodan

1:25 PM - 31 Mar 2016 · Rabanal, Puerto Rico, United States

Tali @talius 31 Mar

#Zika virus #Structure revealed, a critical advance in the development of Treatments & Vaccines

The Purdue Uni...
fb.me/161PWd9tE

Tali @talius Follow

AsterixDB

more engine less trunk

Apache AsterixDB project page:

https://asterixdb.apache.org/