APACHE
GEODE
(INCUBATING)

TAXI TRIP ANALYSIS
(DEBS GRAND-CHALLENGE)
WITH APACHE GEODE

Swapnil Bawaskar
sbawaskar@apache.org
Pivotal

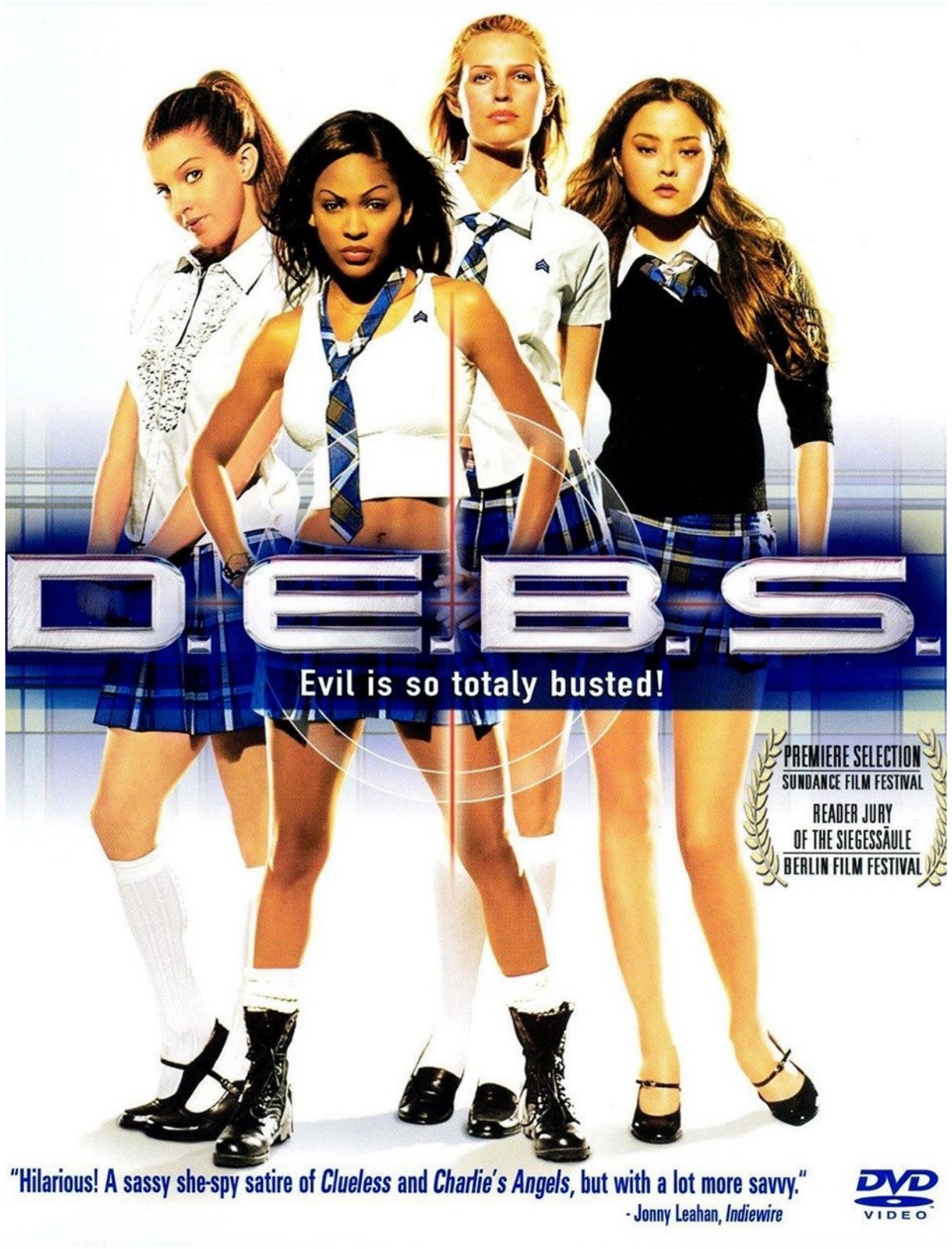
William Markito Oliveira
markito@apache.org
Pivotal

DEBS

- Distributed Event-Based Systems
- Grand challenges (2013, 2014, **2015**, 2016…)
  - Analyze NY Taxi Trip information 2013*
    - 12 GB in size and ~173 million events.
    - Most profitable areas
    - Most frequent routes

* FOIL (The Freedom of Information Law)

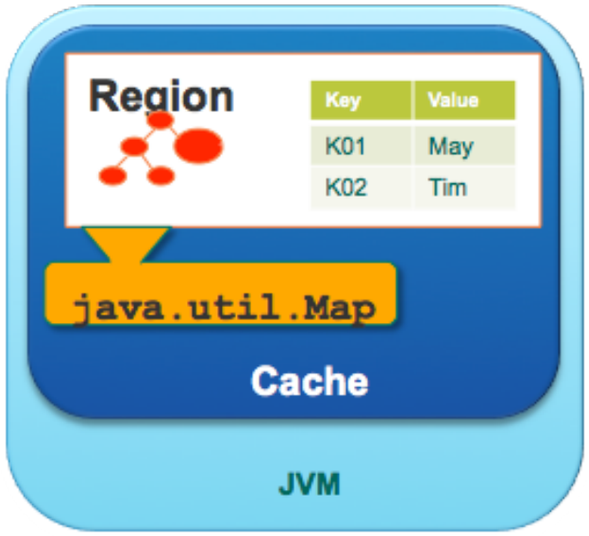
DEBS


D.E.B.S.
Evil is so totaly busted!
PREMIERE SELECTION
SUNDANCE FILM FESTIVAL
READER JURY
OF THE SIEGESSÄULE
BERLIN FILM FESTIVAL
"Hilarious! A sassy she-spy satire of Clueless and Charlie's Angels, but with a lot more savvy."
- Jonny Leahan, Indiewire
DVD VIDEO


RONALDO

BASICS AND TERMINOLOGY

- Cache
  - Configurable through XML, spring or plain Java.
- Region
  - Distributed j.u.Map on steroids (K/V API)
  - Highly available, redundant, persistent
- Member
  - Locator, Server and Client
- OQL - Object Query Language



* Incubating since 2015/May but more than 10 years in development known as GemFire

SOME REFERENCES...

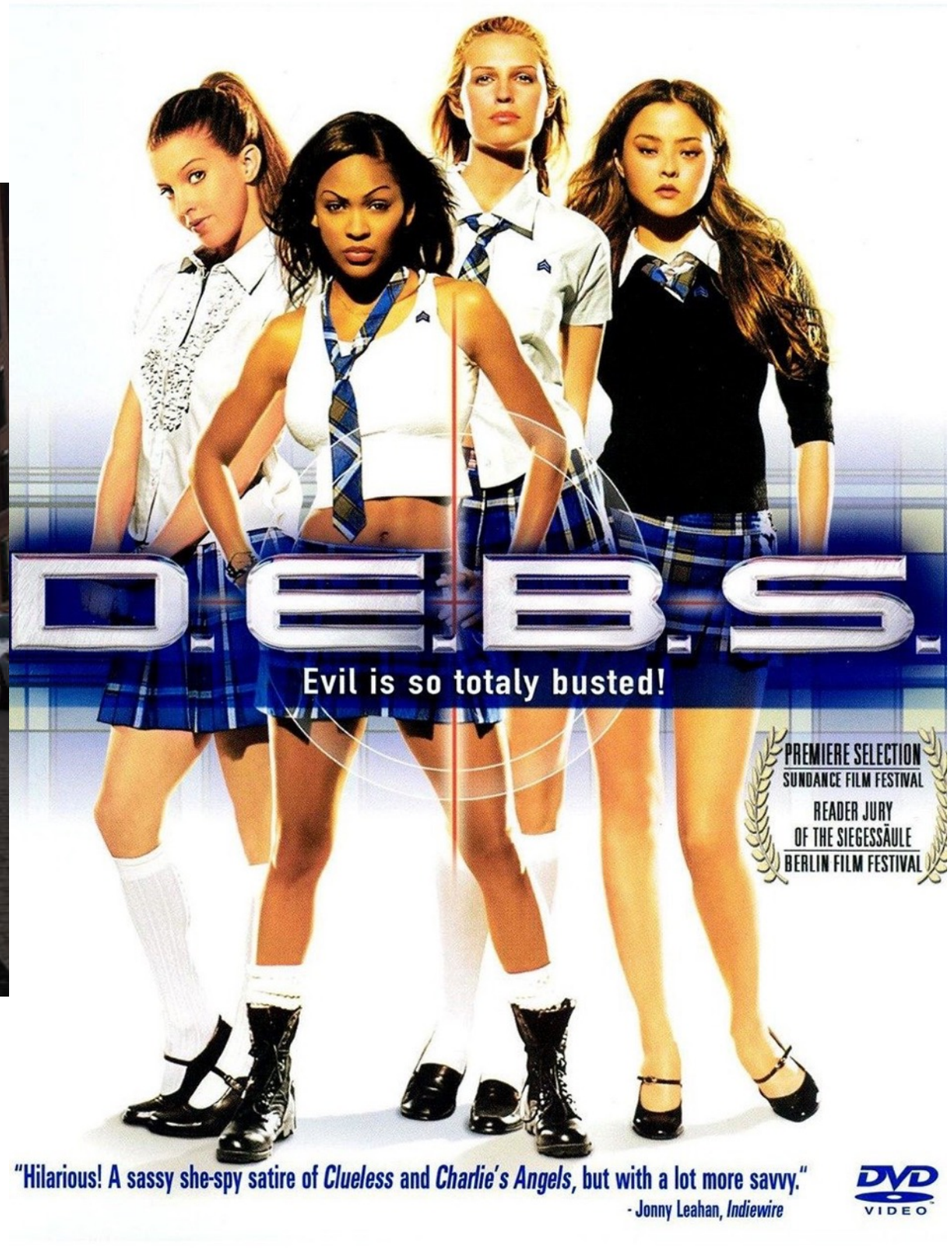
China Railway Corporation

5,700 train stations
4.5 million tickets per day
20 million daily users
1.4 billion page views per day
40,000 visits per second

Indian Railways

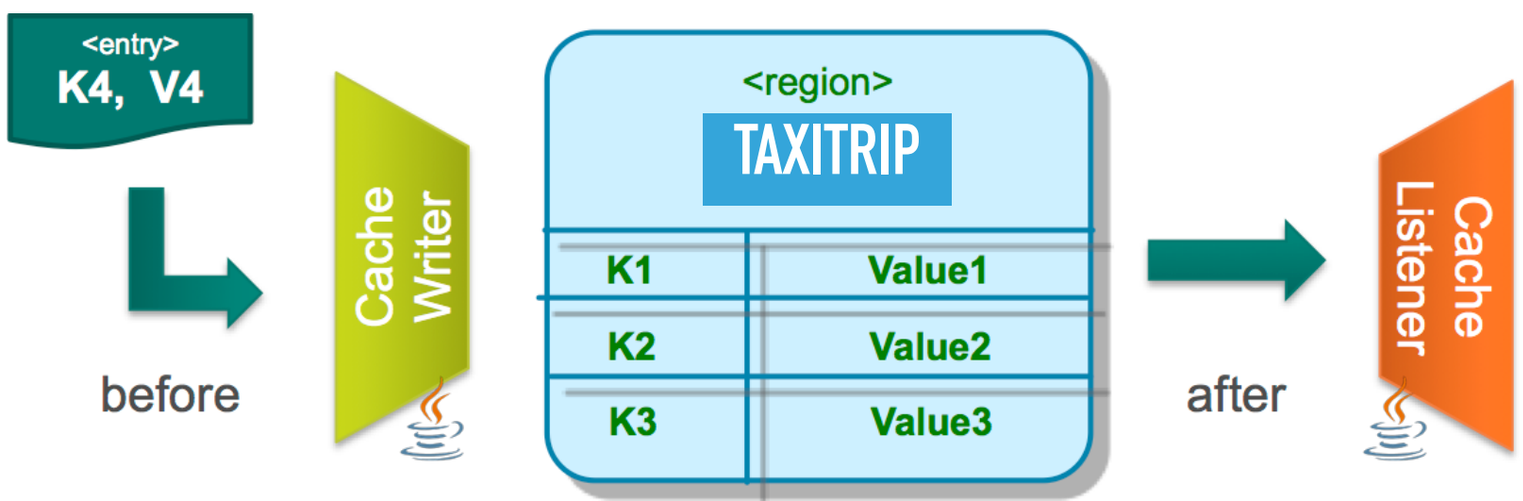
7,000 stations
72,000 miles of track
23 million passengers daily
120,000 concurrent users
10,000 transactions per minute

IMPLEMENTATION



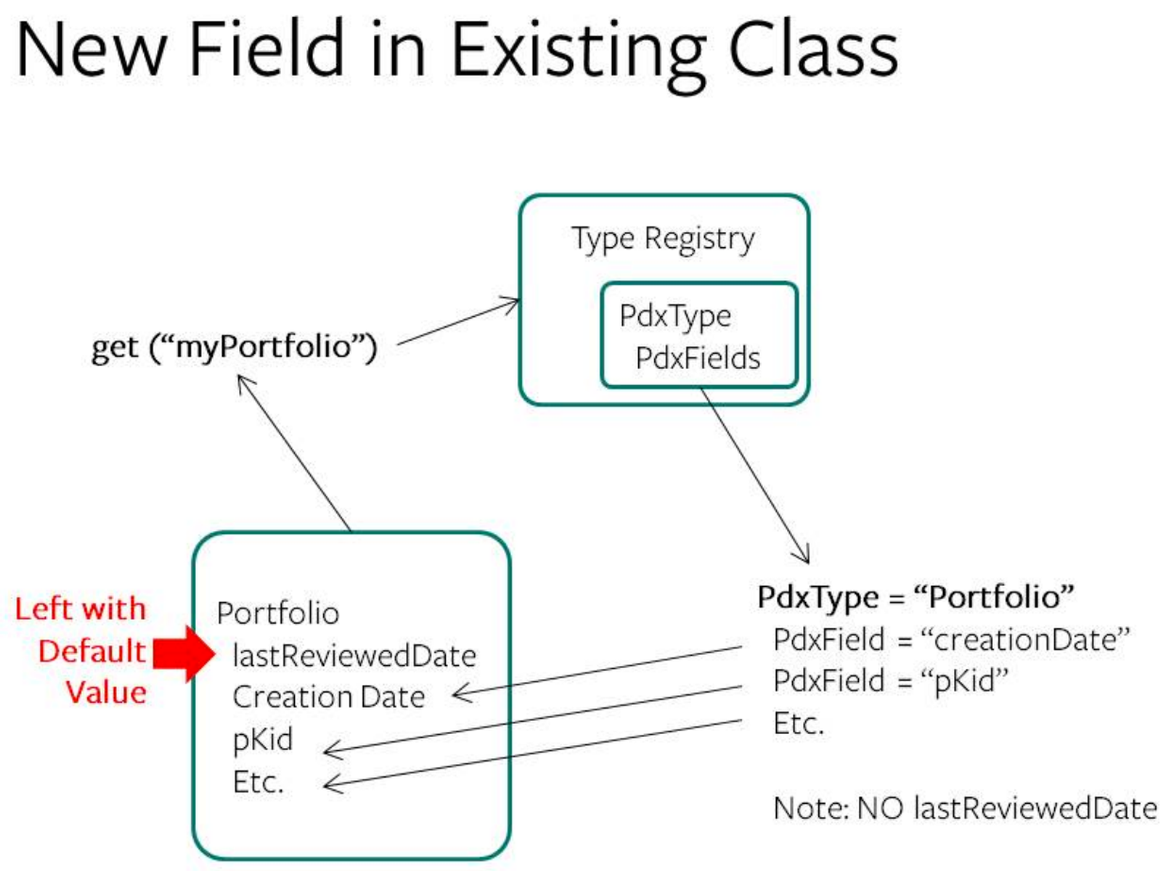
HOW

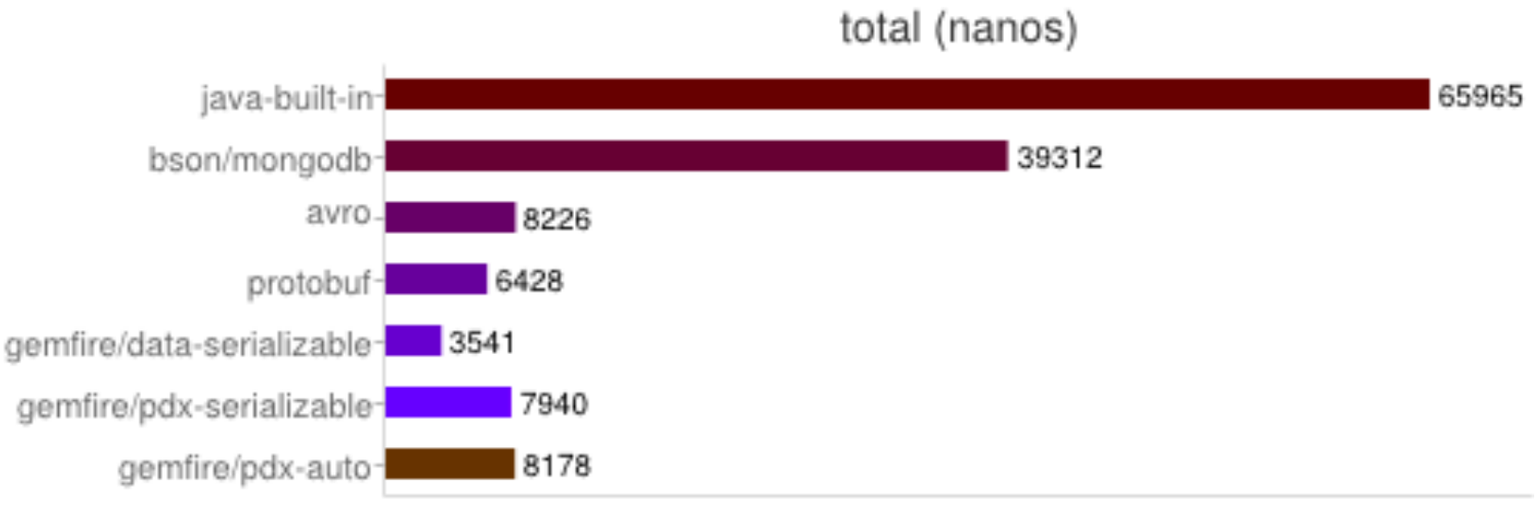
- PDX - (Portable Data eXchange)
  - Compressed, by-field deserialization on demand, etc…
- Functions
  - Distributed Java code with failover (MapReduce like)
    - .onServer, onServers, onRegion (data-aware)
- Callbacks
  - Listener, Writer, AsyncEventListener, Parallel/Serial



HOW

- PDX



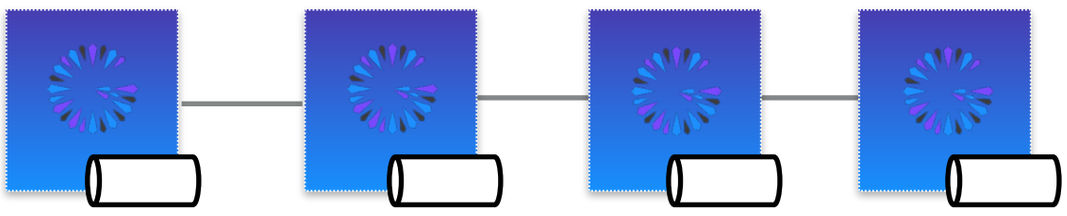


https://blog.pivotal.io/pivotal/products/data-serialization-how-to-run-multiple-big-data-apps-at-once-with-gemfire

HOW

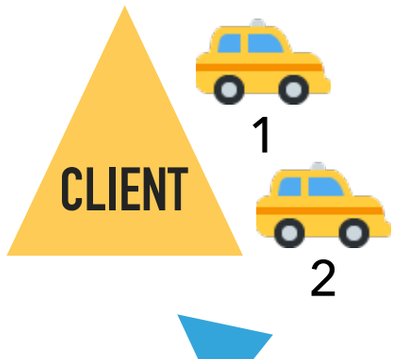
- AsyncEvent Listener
  - Parallel or Serial

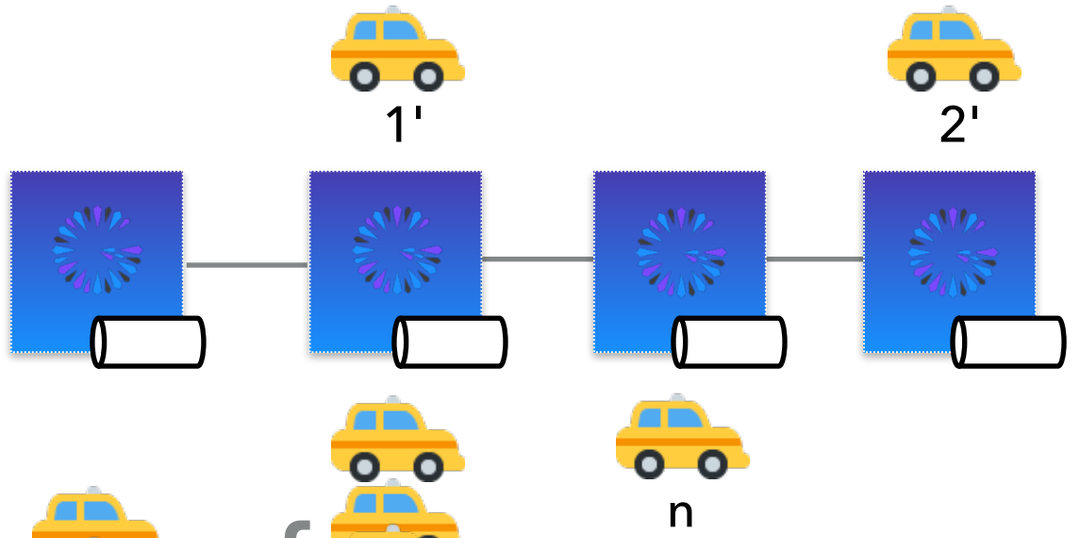
```
public class FrequentRouterListener implements AsyncEventListener, Declarable {
…
  public boolean processEvents(List<AsyncEvent> list)
  {
  …
    // PDX object deserializing single field
    pickupDatetime = (Date) taxiTrip.getField("pickup_datetime");
    …
    // some processing with events
  }

}
```



- Memory
- Threads
- Persistence
- Batch size
- Batch interval

HOW

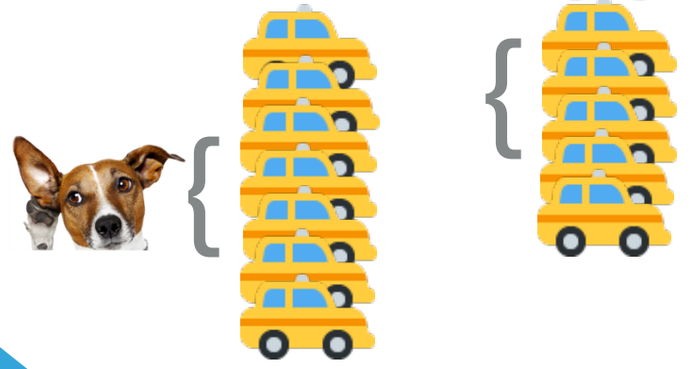




F_ROUTES

Area	Area
1.1	x.y
2.1	x'.y'

CACHING_PROXY



TRIPS

Taxi	Area
1	x.y
2	x'.y'
N	x''.y

Update routes

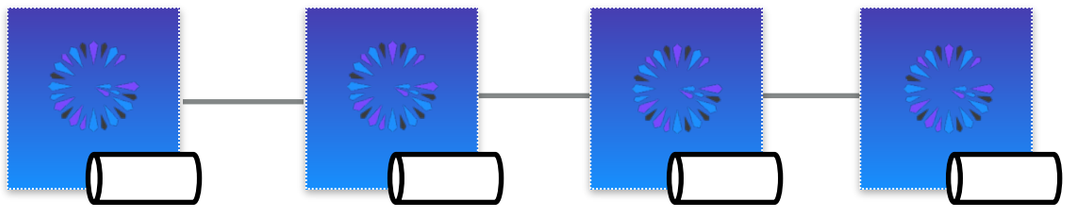
F_ROUTES

Area	Area
1.1	x.y
2.1	x'.y'

```
SELECT AVG(getFarePlusTip()) as avgTotal, pickup_cell.toString()
FROM /TaxiTrip t GROUP BY pickup_cell.toString() ORDER BY avgTotal DESC LIMIT 10"
```

NOT SQL!*

HOW



F_ROUTES

Area	Area
1.1	x.y
2.1	x'.y'

- Evict entries based on entry count (LRU)
- Replicated
- Listener attached

TRIPS

Taxi	Area
1	x.y
2	x'.y'
N	x''.y

- Historical with memory eviction to disk
- Partitioned across nodes
- Async listener with queue

DEMO

JOIN US!

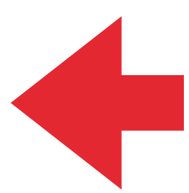
APACHE GEODE

- Mailing lists
  - user-subscribe@geode.incubator.apache.org
  - dev-subscribe@geode.incubator.apache.org
- Events and Virtual Meetup
  - YouTube channel - http://bit.ly/1GZuvcK
  - http://geode.incubator.apache.org/community/

Come talk to us at THE APACHE SOFTWARE FOUNDATION booth and grab a sticker

REFERENCES AND LINKS

- Photos
  - http://www.cosmopolitan.com/sex-love/news/a49615/nyc-sexiest-cab-drivers/
- DEBS Grand Challenge
  - 2015 Challenge
    - debs2015.org/call-grand-challenge.html
  - Data set (12GB)
    - http://chriswhong.com/open-data/foil_nyc_taxi/
- Apache Geode
  - geode.incubator.apache.org
- Implementation
  - https://github.com/markito/debs2015-geode



APACHE GEODE

THANK YOU.

geode.incubator.apache.org

WE'RE HIRING!
Pivotal