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

Swapnil Bawaskar
sbawaskar@apache.org

William Markito Oliveira
markito@apache.org
INTRODUCTION

- Distributed Event-Based Systems
  - 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)
INTRODUCTION

DEBS

“Hilarious! A sassy she-spy satire of Clueless and Charlie’s Angels, but with a lot more savvy.”

- Jonny Lehan, Indiewire
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
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
IMPLEMENTATION

HOW

▸ PDX

New Field in Existing Class

IMPLEMENTATION

HOW

- AsyncEvent Listener
- Parallel or Serial

```java
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
    }
}
```
**IMPLEMENTATION**

**HOW**

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

**CACHING_PROXY**

**F_ROUTES**

<table>
<thead>
<tr>
<th>Area</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>x.y</td>
</tr>
<tr>
<td>2.1</td>
<td>x'.y'</td>
</tr>
</tbody>
</table>

**TRIPS**

<table>
<thead>
<tr>
<th>Taxi</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x.y</td>
</tr>
<tr>
<td>2</td>
<td>x'.y'</td>
</tr>
</tbody>
</table>
| N    | x```.y```

**Update routes**
**IMPLEMENTATION**

**HOW**

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

---

**F_ROUTES**

<table>
<thead>
<tr>
<th>Area</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>x.y</td>
</tr>
<tr>
<td>2.1</td>
<td>x'.y'</td>
</tr>
</tbody>
</table>

**TRIPS**

<table>
<thead>
<tr>
<th>Taxi</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x.y</td>
</tr>
<tr>
<td>2</td>
<td>x'.y'</td>
</tr>
<tr>
<td>N</td>
<td>x''.y</td>
</tr>
</tbody>
</table>

- Historical with memory eviction to disk
- Partitioned across nodes
- Async listener with queue
DEMO
JOIN US!

- Mailing lists
  - user-subscribe@geode.incubator.apache.org
  - dev-subscribe@geode.incubator.apache.org
- Events and Virtual Meetup
  - http://geode.incubator.apache.org/community/

Come talk to us at booth and grab a sticker
REFERENCES AND LINKS

- Photos

- 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
THANK YOU.

geode.incubator.apache.org