UC Berkeley's and Caltrans' new cloud-based data hub

Qijian Gan
Postdoctoral Researcher
PATH
University of California, Berkeley
qgan@berkeley.edu

For the TRB Workshop on Big Data Applications and Methods in Transportation

January 7th, 2018
Outline

- Introduction
- Architecture
- Application Examples
- Conclusion
Introduction
Why ICM (Integrated Corridor Management)?

- Traffic is getting worse and we cannot only build more roads
- To improve network performance, it is time to consider corridor-wide management that operates the network in a more “coordinated” way
- TSM&O (Transportation Systems Management and Operations)
- Caltrans created a new statewide program: **Connected Corridors**
- **Pilot in LA**: Traffic in LA is one of the worst in the U.S. *(Source: TTI 2015 Urban Mobility Scorecard)*
What is Connected Corridors?

- A statewide program – [https://connected-corridors.berkeley.edu/home](https://connected-corridors.berkeley.edu/home)
- The integration of multiple components into a traffic management system
  - Not a simple piece of technology
  - A total entity made up of people, organizations, hardware, and software
A significant number of daily traffic incidents
- Heavily instrumented: good sensing coverage
- Cooperation of cities and the county
- Arterial network has some capacity to accommodate additional traffic
I-210 Pilot ICM: Connected Systems
The PATH Connected Corridors Team

**Senior Leadership**

Alex Bayen  
Thomas West  
Joe Butler

**Faculty**

Alex Bayen  
Adib Kanafani  
Alex Skabardonis

**Research Staff**

Francois Dion  
Anthony Patire  
Gabriel Gomes  
Qijian Gan

**Program Staff**

Brian Peterson  
Shivani Bongani  
Jeny Govindan  
Gary Gremaux  
Michelle Harrington  
Tom Kuhn  
Cindy Li  
Greg Merritt  
Sean Morris  
Jessica Rojas  
Laman Sadaghiani  
Nathaniel Titterton

**Caltrans Partners**

Nick Compin  
Raju Porandla  
Allen Chen  
Farid Nowshiravan
Technical Architecture
High Level Data Flow

Data Hub

Amazon Web Services

Data Sources/TMCs

Decision Support System (DSS)

Corridor Management System (CMS)

Control Targets/TMCs
Design Challenges

- Flexibility & Scalability ⟷ Amazon Web Services
- Different data types ⟷ Different storage implementations
- Data quality ⟷ Data processing and transformation
- Computation ⟷ SPARK with Machine Learning
- Data movement ⟷ Kafka/ActiveMQ messaging
- Process control & Orchestration ⟷ Netflix Conductor for work flows
Connected Corridors ICM Architecture
Data Processing Strategies

- **Real Time Data Streams**
  - Data from field sensors (freeway & arterial), intersection signals, and probes
    - High frequency & high volume & low relational content
  - Current solution:
    - Kafka + SPARK + Cassandra

- **Heterogeneous Sources**
  - Data for Intersection Signal Inventory/State
    - Low frequency & with different subtleties & more relational content
  - Current solution:
    - Mongo + Kafka/ActiveMQ + Postgres

- **Homogeneous Sources**
  - Data for Ramp Meter Inventory/State
    - Low frequency & with a common format & more relational content
  - Current solution:
    - Kafka/ActiveMQ + Postgres
## Major Data Sources

<table>
<thead>
<tr>
<th>Source</th>
<th>Information Type</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pasadena</td>
<td>Intersection signal</td>
<td>Pasadena TMC</td>
</tr>
<tr>
<td>Duarte</td>
<td>Intersection signal</td>
<td>County TMC</td>
</tr>
<tr>
<td>Monrovia</td>
<td>Intersection signal</td>
<td>County TMC</td>
</tr>
<tr>
<td>Arcadia</td>
<td>Intersection signal</td>
<td>Arcadia TMC</td>
</tr>
<tr>
<td>LA County</td>
<td>Intersection signal</td>
<td>County TMC</td>
</tr>
<tr>
<td>Caltrans FW Traffic</td>
<td>Loop sensing</td>
<td>Caltrans ATMS</td>
</tr>
<tr>
<td>Caltrans FW Ramps</td>
<td>Ramp meters</td>
<td>Caltrans ATMS</td>
</tr>
<tr>
<td>Caltrans FW CMS</td>
<td>DMS</td>
<td>Caltrans ATMS</td>
</tr>
<tr>
<td>Caltrans Intersections</td>
<td>Intersection signal</td>
<td>TSMSS</td>
</tr>
<tr>
<td>Caltrans Video</td>
<td>Video</td>
<td>via RIITS</td>
</tr>
<tr>
<td>Caltrans FW Lane closure</td>
<td>Lane status</td>
<td>LCS</td>
</tr>
<tr>
<td>Caltrans incident</td>
<td>Incident</td>
<td>Caltrans ATMS</td>
</tr>
<tr>
<td>210 LCS</td>
<td>Lane status</td>
<td>High speed rail system</td>
</tr>
<tr>
<td>RIITTS Environmental sensing</td>
<td>Environmental</td>
<td>RIITTS</td>
</tr>
<tr>
<td>RIITTS Transit</td>
<td>Transit</td>
<td>RIITTS</td>
</tr>
<tr>
<td>RIITTS Video</td>
<td>Video</td>
<td>RIITTS</td>
</tr>
<tr>
<td>Gold line transit</td>
<td>Transit</td>
<td>NextBus</td>
</tr>
<tr>
<td>511 (Out only)</td>
<td>Response plan information</td>
<td></td>
</tr>
<tr>
<td>Bluetooth traffic</td>
<td>Travel time</td>
<td>County TMC</td>
</tr>
</tbody>
</table>
Real-time Data Streams

Data Example: Arterial Sensor Data

<table>
<thead>
<tr>
<th>DetectorID</th>
<th>Date</th>
<th>Time</th>
<th>State</th>
<th>Speed</th>
<th>Occupancy</th>
<th>Volume</th>
<th>AvgSpeed</th>
<th>AvgOccupancy</th>
<th>AvgVolume</th>
</tr>
</thead>
<tbody>
<tr>
<td>307502</td>
<td>20170100</td>
<td>48541</td>
<td>OPERATIONAL</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>50.00</td>
<td>1.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170104</td>
<td>34715</td>
<td>OPERATIONAL</td>
<td>15.00</td>
<td>6.00</td>
<td>240.00</td>
<td>21.00</td>
<td>7.00</td>
<td>325.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170105</td>
<td>71725</td>
<td>OPERATIONAL</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>50.00</td>
<td>1.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170109</td>
<td>75063</td>
<td>OPERATIONAL</td>
<td>24.00</td>
<td>10.00</td>
<td>600.00</td>
<td>26.00</td>
<td>9.00</td>
<td>522.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170122</td>
<td>82462</td>
<td>OPERATIONAL</td>
<td>16.00</td>
<td>7.00</td>
<td>260.00</td>
<td>21.00</td>
<td>6.00</td>
<td>267.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170123</td>
<td>22357</td>
<td>OPERATIONAL</td>
<td>23.00</td>
<td>4.00</td>
<td>240.00</td>
<td>15.00</td>
<td>3.00</td>
<td>137.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170123</td>
<td>47839</td>
<td>OPERATIONAL</td>
<td>24.00</td>
<td>8.00</td>
<td>480.00</td>
<td>23.00</td>
<td>13.00</td>
<td>658.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170127</td>
<td>28596</td>
<td>OPERATIONAL</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170128</td>
<td>54630</td>
<td>OPERATIONAL</td>
<td>0.00</td>
<td>97.00</td>
<td>30.00</td>
<td>1.00</td>
<td>96.00</td>
<td>52.00</td>
</tr>
<tr>
<td>307502</td>
<td>20170129</td>
<td>1510</td>
<td>OPERATIONAL</td>
<td>15.00</td>
<td>24.00</td>
<td>1200.00</td>
<td>19.00</td>
<td>22.00</td>
<td>1043.00</td>
</tr>
</tbody>
</table>
Heterogeneous Sources

Data Example: Intersection Signal Inventory

<table>
<thead>
<tr>
<th>OrgID</th>
<th>IntersectionID</th>
<th>Date</th>
<th>Time</th>
<th>SignalType</th>
<th>Description</th>
<th>MainStreet</th>
<th>CrossStreet</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:1</td>
<td>3075</td>
<td>20170626</td>
<td>42943</td>
<td>170 LACO IV</td>
<td>Foothill / Second</td>
<td>Foothild Blvd</td>
<td>Second Ave</td>
<td>34.151051</td>
<td>-118.025402</td>
</tr>
<tr>
<td>5:1</td>
<td>3076</td>
<td>20170626</td>
<td>42944</td>
<td>170 LACO IV</td>
<td>Foothill / First</td>
<td>Foothild Blvd</td>
<td>First Ave</td>
<td>34.151014</td>
<td>-118.026533</td>
</tr>
<tr>
<td>5:1</td>
<td>3077</td>
<td>20170626</td>
<td>42945</td>
<td>2070 D4 I SL</td>
<td>Foothill / Santa Anita</td>
<td>Foothild Blvd</td>
<td>Santa Anita Ave</td>
<td>34.150945</td>
<td>-118.035164</td>
</tr>
<tr>
<td>5:1</td>
<td>3078</td>
<td>20170626</td>
<td>42946</td>
<td>170 LACO IV</td>
<td>Foothill / Baldwin</td>
<td>Foothild Blvd</td>
<td>Baldwin Ave</td>
<td>34.150958</td>
<td>-118.030237</td>
</tr>
<tr>
<td>5:1</td>
<td>3079</td>
<td>20170626</td>
<td>42948</td>
<td>170 LACO IV</td>
<td>Colorado / Baldwin</td>
<td>Colorado St</td>
<td>Baldwin Ave</td>
<td>34.148526</td>
<td>-118.030245</td>
</tr>
<tr>
<td>5:1</td>
<td>3080</td>
<td>20170626</td>
<td>42887</td>
<td>2070 D4 I SL</td>
<td>Baldwin / Gate 7</td>
<td>Baldwin Ave</td>
<td>Gate 7</td>
<td>34.142715</td>
<td>-118.051142</td>
</tr>
<tr>
<td>5:1</td>
<td>3081</td>
<td>20170626</td>
<td>42888</td>
<td>2070 D4 I SL</td>
<td>Baldwin / Gate 8</td>
<td>Baldwin Ave</td>
<td>Gate 8</td>
<td>34.138967</td>
<td>-118.051011</td>
</tr>
<tr>
<td>5:1</td>
<td>3082</td>
<td>20170626</td>
<td>42898</td>
<td>2070 D4 I SL</td>
<td>Baldwin / Gate 9</td>
<td>Baldwin Ave</td>
<td>Gate 9</td>
<td>34.136753</td>
<td>-118.054136</td>
</tr>
<tr>
<td>5:1</td>
<td>3091</td>
<td>20170626</td>
<td>42890</td>
<td>2070 D4 I SL</td>
<td>Baldwin / Gate 10</td>
<td>Baldwin Ave</td>
<td>Gate 10</td>
<td>34.134092</td>
<td>-118.054322</td>
</tr>
<tr>
<td>5:1</td>
<td>3092</td>
<td>20170626</td>
<td>42891</td>
<td>2070 D4 I SL</td>
<td>Huntington / Baldwin</td>
<td>Huntington Blvd</td>
<td>Baldwin</td>
<td>34.131693</td>
<td>-118.054503</td>
</tr>
<tr>
<td>5:1</td>
<td>3093</td>
<td>20170626</td>
<td>42892</td>
<td>2070 D4 I SL</td>
<td>Huntington / Gate 1</td>
<td>Huntington Blvd</td>
<td>Gate 1</td>
<td>34.131808</td>
<td>-118.051978</td>
</tr>
<tr>
<td>5:1</td>
<td>3094</td>
<td>20170626</td>
<td>42893</td>
<td>2070 D4 I SL</td>
<td>Huntington / La Cadena</td>
<td>Huntington Blvd</td>
<td>La Cadena</td>
<td>34.131757</td>
<td>-118.046656</td>
</tr>
<tr>
<td>5:1</td>
<td>3095</td>
<td>20170626</td>
<td>42894</td>
<td>2070 D4 I SL</td>
<td>Huntington / Michilinda</td>
<td>Huntington Dr</td>
<td>Michilinda Ave</td>
<td>34.130806</td>
<td>-118.067429</td>
</tr>
<tr>
<td>5:1</td>
<td>3096</td>
<td>20170626</td>
<td>42895</td>
<td>2070 D4 I SL</td>
<td>Huntington / Sunset</td>
<td>Huntington Dr</td>
<td>Sunset Blvd</td>
<td>34.130945</td>
<td>-118.065124</td>
</tr>
</tbody>
</table>
Homogeneous Sources

Ramp Meter Inventory/State → Ramp Meter Reader → Ramp Meter Processor → ActiveMQ/Kafka → Postgres/PostGIS → DSS

Ramp Meter Reader → ActiveMQ/Kafka

Ramp Meter Processor → ActiveMQ/Kafka
Orchestration – Command Gateway

- **Inside the data hub: Independent services connected by messaging**
  - Each service has specific functions without knowledge of other services
  - Messaging connects services for data flows and control flows

- **External Interface: Independent pipelines with workflow and data flow control**
  - Each data pipeline has no knowledge of the other pipelines or how to communicate with the other pipelines
  - The knowledge of workflow and message routing stays in a central place - Command Gateway
Application Examples
Machine Learning – Flow Prediction

Three models from Mlib currently deployed in the Data Hub:
- Gradient boosted tree (Best Performance)
- Random forest
- Linear autoregressive model

Flow prediction at a mainline sensor

Prediction Horizon

Prediction Error

Data required for prediction
Up-to-date model parameters
Prediction outputs
Freeway Traffic Estimation

- Real-time PeMS Data
- Freeway Network

Cell Transmission Model (CTM)

Ensemble Kalman Filter (EnKF)

Estimated Traffic States (Flow, Density, Speed)
Arterial Queue Estimation

- Arterial Detector Data
- Arterial Signal Phasing Data
- Intersection Road Geometry

Data Hub ➔ Arterial Traffic State Estimation ➔ Estimated Traffic Queues (Left-Turn, Through, Right-Turn)

• Left-Turn Detector
• Advance Detector 1
• Advance Detector 2

Resulting Queue Estimates
Arterial Data Quality and Detector Health Analysis

- Data quality and detector health analysis (from intersection-level to network-level)
- Will extend to the analysis of Traffic Signal Performance in the near future
Conclusion
UC Berkeley's and Caltrans’ Data Hub

- **The cloud based data hub is:**
  - Designed for Efficiency, Reliability, and Scalability.
  - A new paradigm for managing transportation big data
  - Playing a key role in the Corridor Management System
  - Planned for deployment in late 2018 or early 2019

- **We (Caltrans & PATH) are planning to open source the software and are happy to discuss our design**
  - Research/Project Collaboration
    - Prof. Alex Bayen (bayen@berkeley.edu)
    - Joe Butler (joebutler@path.berkeley.edu)
  - Data Hub
    - Brian Peterson (brian.peterson@berkeley.edu)
  - Model Development
    - Qijian Gan (qgan@berkeley.edu)
Thank you!
Supporting Slides
/Q&A Support
## Technology Stack (1/2)

<table>
<thead>
<tr>
<th>Technology</th>
<th>Purpose</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Java 7/8</td>
<td>Primary server-side programming language/framework</td>
<td>Broadly understood, easy to find resources, lots of experience/tools</td>
<td>Can be complex</td>
</tr>
<tr>
<td>Cassandra (OS/Commercial) 3.10</td>
<td>High volume, real time time-series data (sensing/probe)</td>
<td>Very fast with large data volumes, highly scalable, fault tolerant</td>
<td>No ad-hoc querying, limited talent/resources</td>
</tr>
<tr>
<td>MongoDB (OS/Commercial) 3.4.4</td>
<td>Transformation of complex relational structures</td>
<td>Document storage (schema-less), very fast querying</td>
<td>Limited talent/resources</td>
</tr>
<tr>
<td>Drools Community v.6.5.0</td>
<td>Rules engine</td>
<td>Widely used java rules engine, large production base</td>
<td>Community version has limited support</td>
</tr>
<tr>
<td>Postgres 9.6.2</td>
<td>Relational data store</td>
<td>Large installed base, used within Caltrans already, easy to find resources, PostGIS for geospatial, AWS hosted service</td>
<td>Not as scalable for extremely large data sets</td>
</tr>
<tr>
<td>Spark 2.1.0</td>
<td>High speed analytics and stream processing (sensor/probe), machine learning platform</td>
<td>Exceptionally fast and scalable processing, AWS hosted service</td>
<td>Limited talent/resources</td>
</tr>
<tr>
<td>Tomcat 8.5.15</td>
<td>WS</td>
<td>Large installed base</td>
<td></td>
</tr>
</tbody>
</table>
## Technology Stack (2/2)

<table>
<thead>
<tr>
<th>Technology</th>
<th>Purpose</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>ActiveMQ 5.14.5</td>
<td>Decoupling mechanism, control messaging, status messaging, large structure data messaging</td>
<td>Significant installed base, broadly understood, capable of large messages</td>
<td>Not the fastest gun in town, not as easily scalable</td>
</tr>
<tr>
<td>Kafka 0.10.20</td>
<td>High speed, high volume data messaging</td>
<td>Built for speed, message persistence, scalable, fault tolerant</td>
<td>Reputation for being temperamental, limited to smaller message sizes, limited talent/resources</td>
</tr>
<tr>
<td>Graylog</td>
<td>System Logging</td>
<td>Simple, large installed base</td>
<td></td>
</tr>
<tr>
<td>Camel 2.18.4</td>
<td>Data hub – CMS/DSS interface and switchboard, protocol transformations</td>
<td>Significant installed base, broadly understood</td>
<td></td>
</tr>
<tr>
<td>Conductor 1.8.0</td>
<td>Data pipeline and DSS/CMS/DH command orchestration and workflow management</td>
<td>Extensive production experience at very high scale (Netflix), flexible</td>
<td></td>
</tr>
</tbody>
</table>
# Primary AWS Services

<table>
<thead>
<tr>
<th>Technology</th>
<th>Purpose</th>
<th>Key uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC2</td>
<td>Server processing on demand</td>
<td>Estimation, Prediction, data processing, Persistence workers, Cassandra, MongoDB, other custom workers, messaging, logging</td>
</tr>
<tr>
<td>RDS</td>
<td>Postgres w/PostGIS</td>
<td>Modeling data store (models, corridor asset model element information)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data hub relational store (corridor asset post transformation)</td>
</tr>
<tr>
<td>S3</td>
<td>Storage</td>
<td>Stateful processing</td>
</tr>
<tr>
<td>Security Groups/ VPC/IAM</td>
<td>Cloud/network isolation/identity &amp; access management</td>
<td>Networking/Security/Cloud access</td>
</tr>
<tr>
<td>EMR</td>
<td>Hosted Spark</td>
<td>Analytics, data quality, machine learning</td>
</tr>
<tr>
<td>Cloud Init, Cloud Formation</td>
<td>Deployment</td>
<td>Instance automation, cloud initialization and maintenance</td>
</tr>
<tr>
<td>CloudWatch, CloudTrail</td>
<td>Monitoring</td>
<td>Monitoring</td>
</tr>
<tr>
<td>Key Management Service</td>
<td>Key Management</td>
<td>Security, encryption</td>
</tr>
</tbody>
</table>
Orchestration - Overview
Decision Support System – Design Detail
Aimsun Modeling, Calibration, & Prediction

- The Aimsun Model
  - ~1000 lane miles of road, ~5000 traffic detectors, 459 signalized intersections and control plans, 45 freeway ramp meters, Metro gold line and all bus routes

- Data Inputs
  - Current
    - 2008 SCAG data, observed flow counts from the field, signal timing plans, ramp metering plans, etc.
  - In the near future
    - Predicted demands, Estimated traffic states, Response plans, etc.
Development of Response Plans & Rules Engine

- ~100 alternate arterial routes have been identified
- ~50 message signs to be installed

Response to a given incident may include 1 to 3 alternate routes from the “menu” of ~300 preliminary routes

Factors affecting choice
- Location of incident
- Prevailing traffic conditions on freeway and arterials
- Ability of route to provide effective relief
- Local defined constraints