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

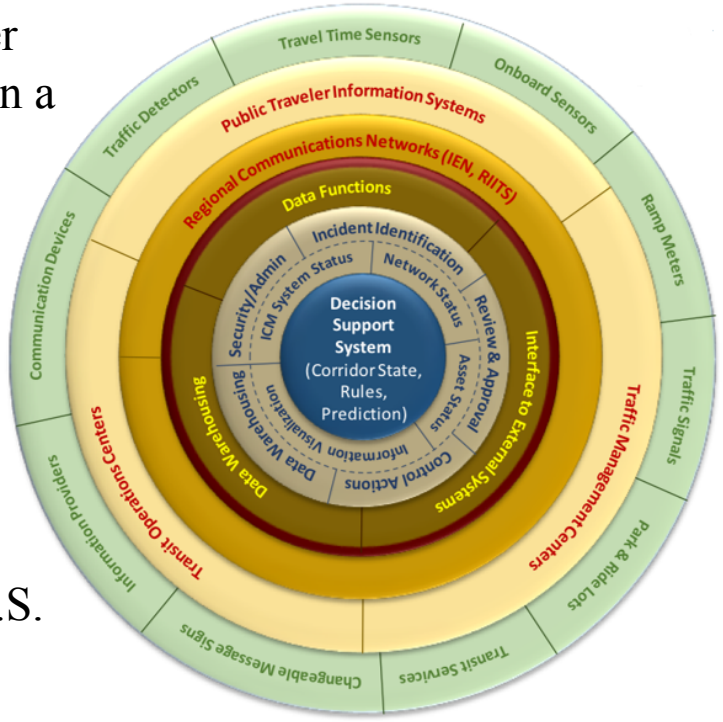
- **Introduction**
- **Architecture**
- **Application Examples**
- **Conclusion**

3

Introduction

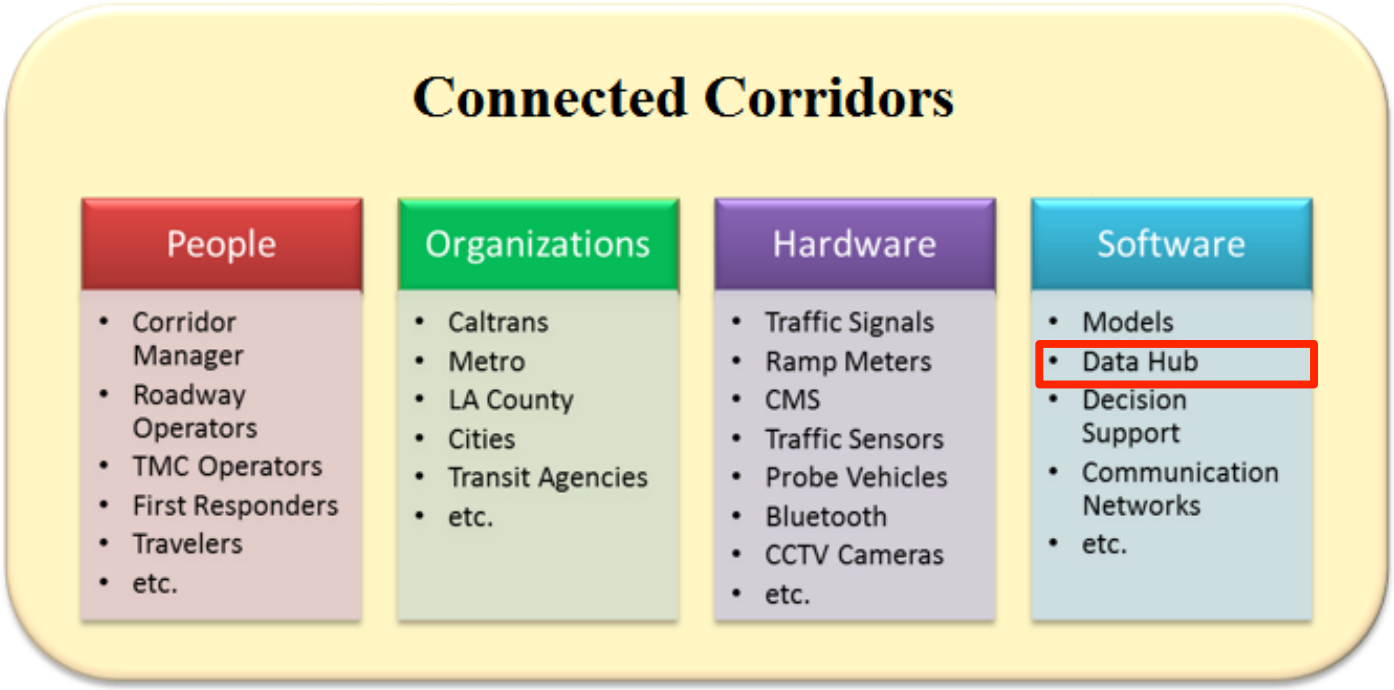
Why ICM (Integrated Corridor Management)?

- 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*)



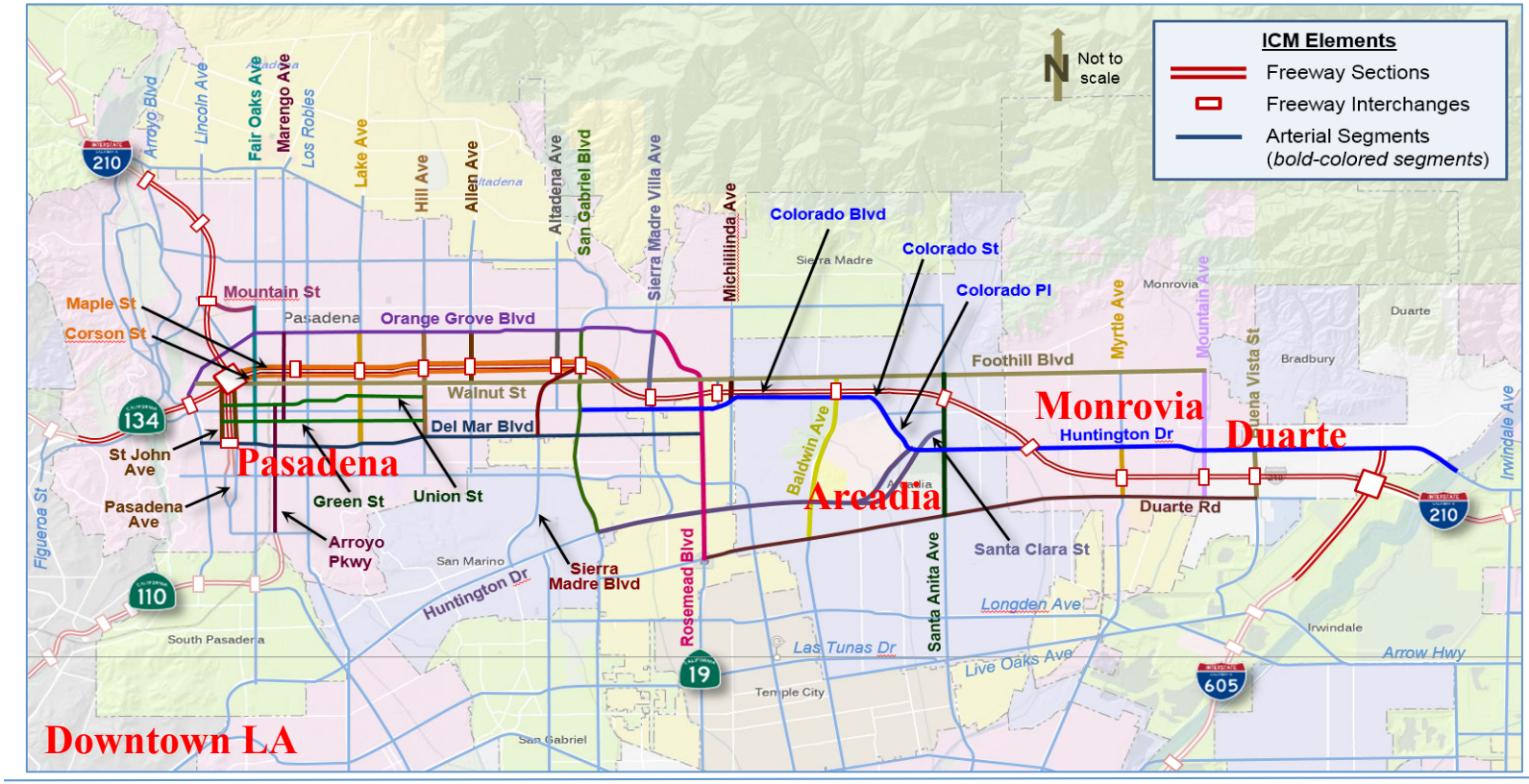
What is Connected Corridors?

- A statewide program – https://connected-corridors.berkeley.edu/home
- 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



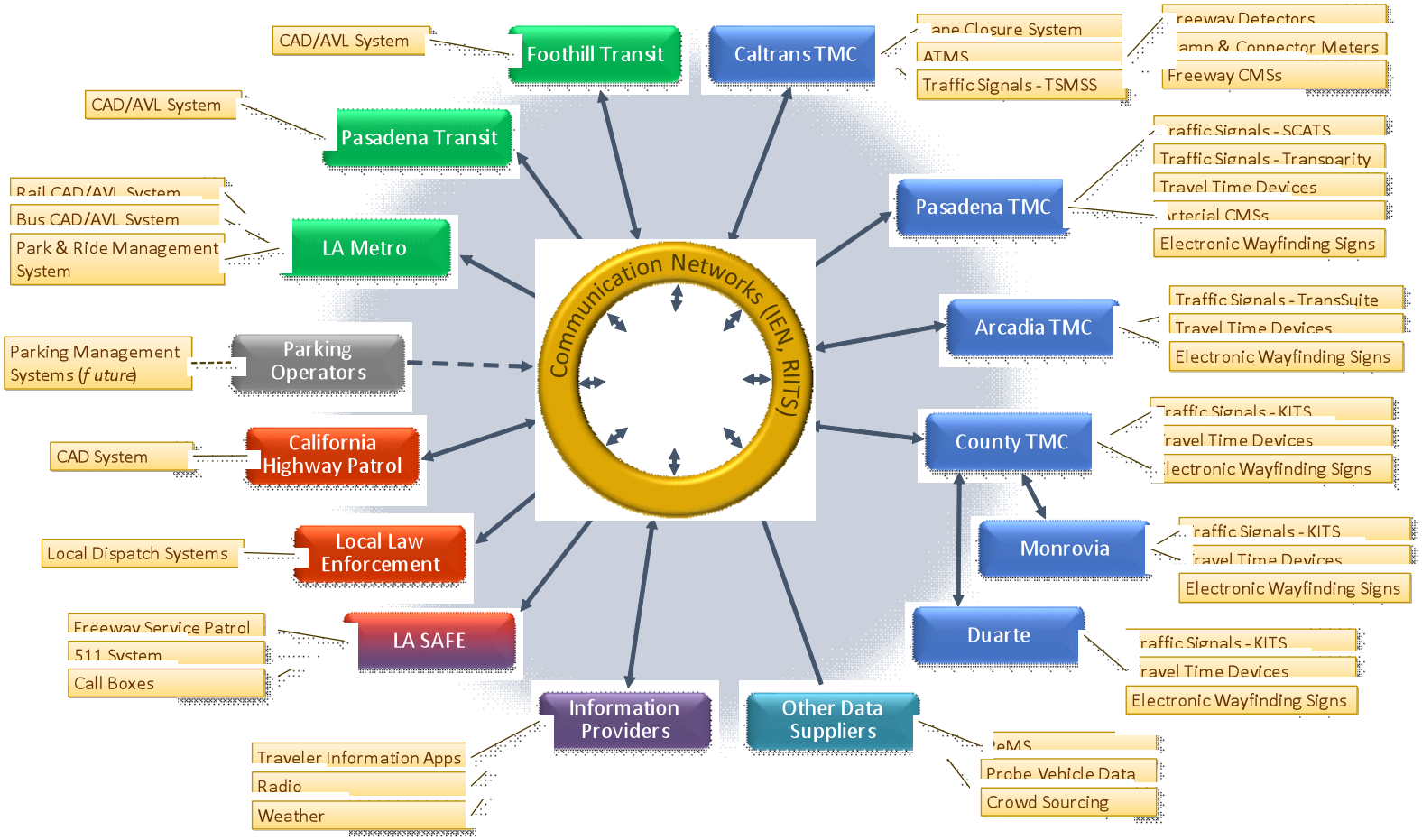
Connected Corridors Pilot - The I-210 in LA

- 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



Caltrans Metro

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

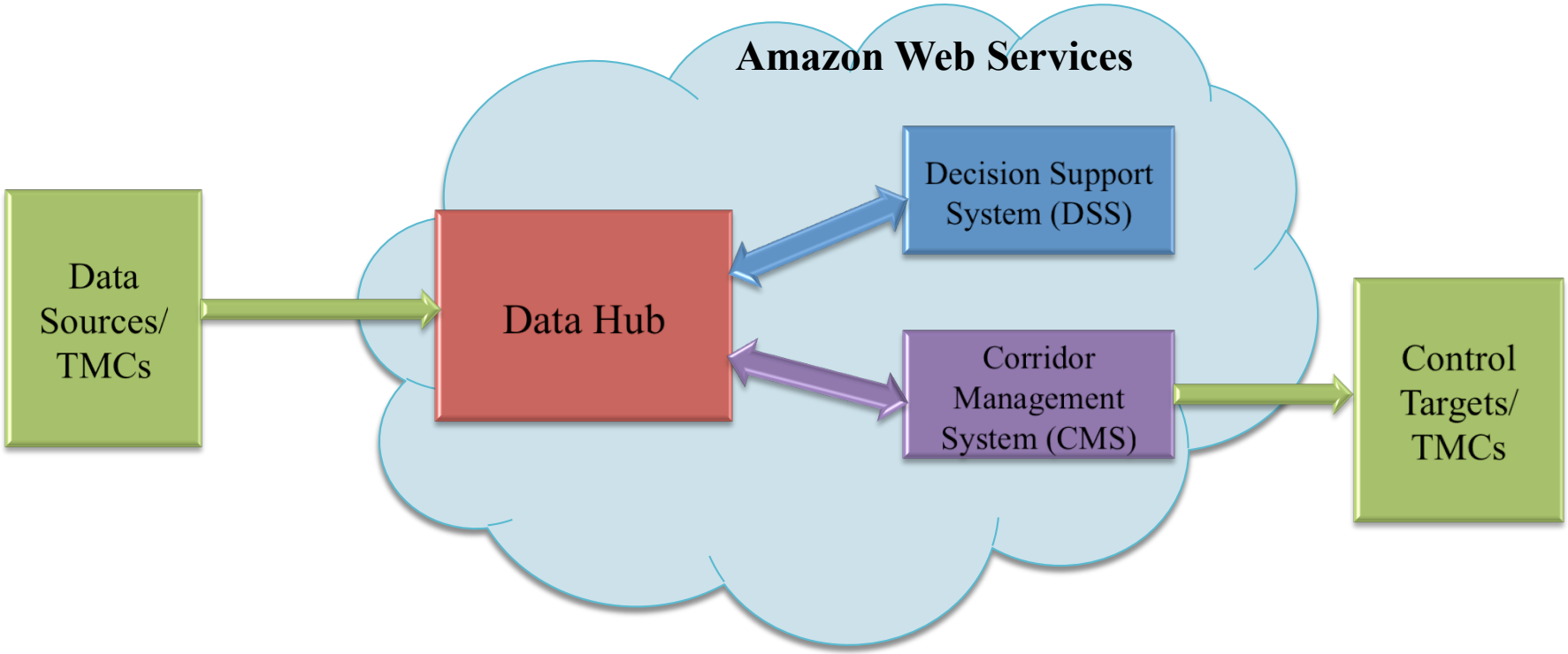

Farid Nowshiravan

Caltrans Metro County of Los Angeles City of Pasadena City of Arcadia City of Monrovia City of Duarte Association of Governments SGVCOG Foothill Transit California PATH

9

Technical Architecture

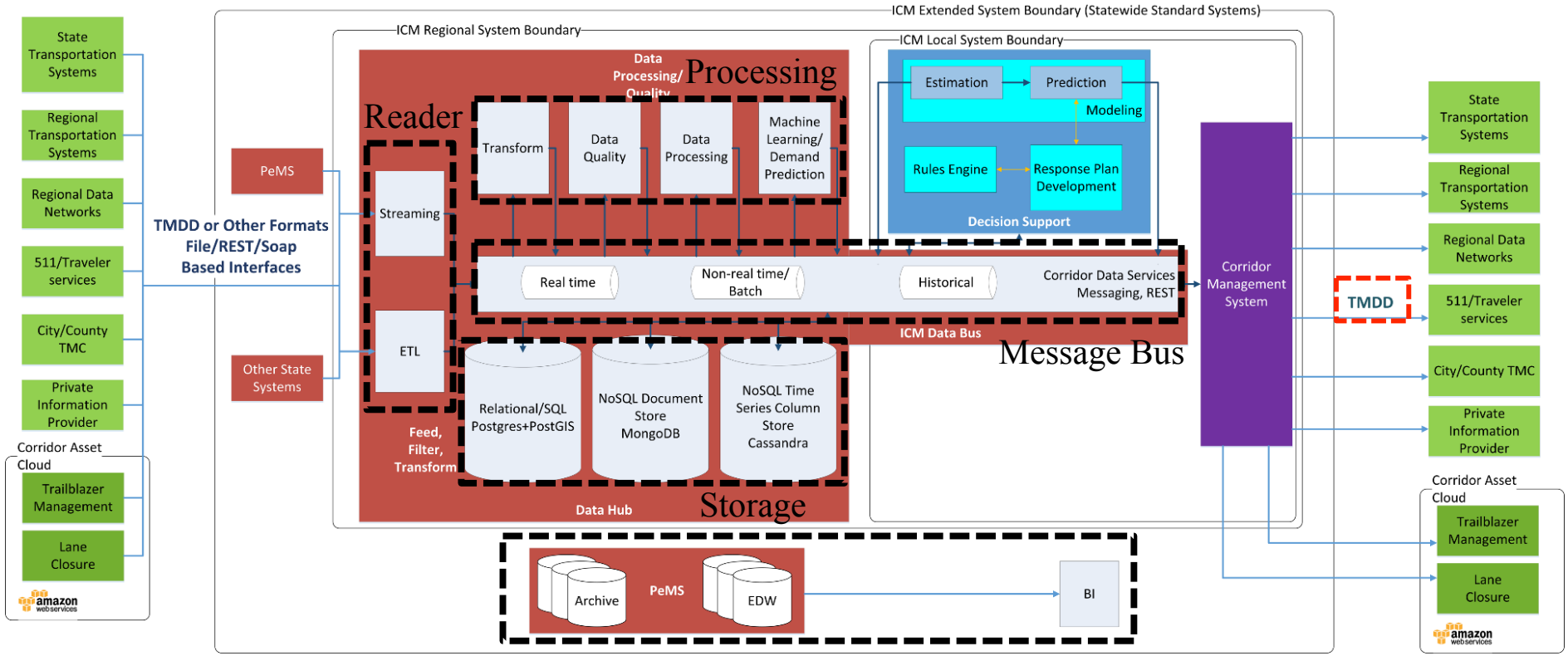
High Level Data Flow



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



Long-term Storage and Analysis

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

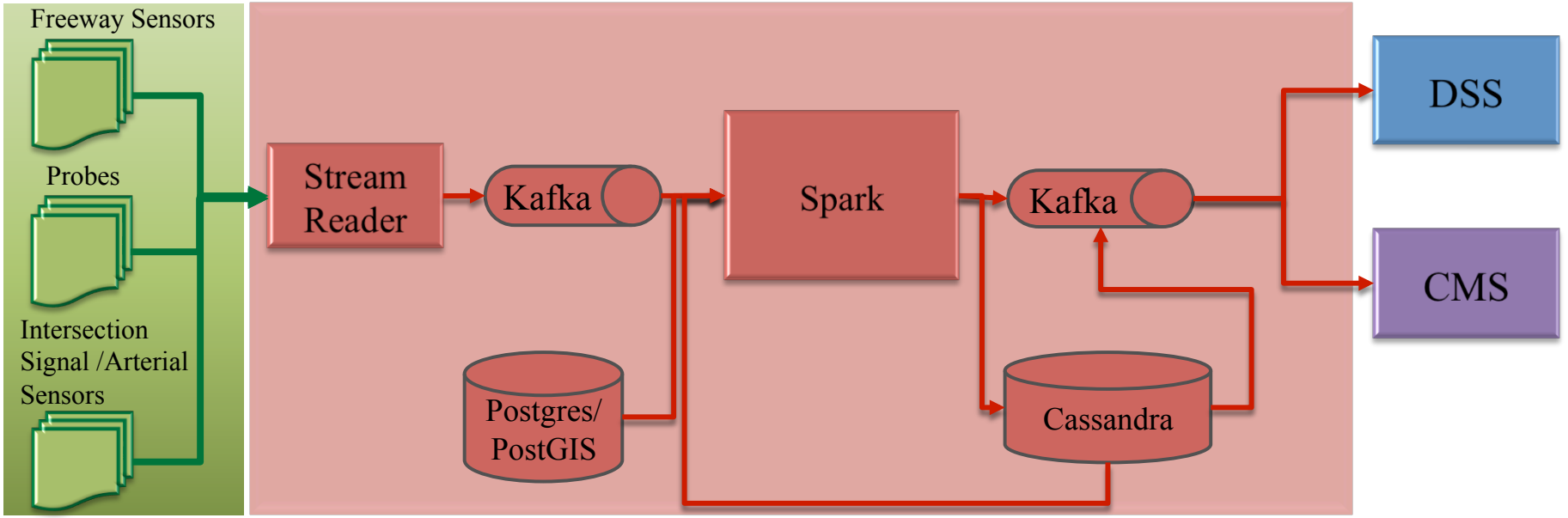
- **Heterogeneous Sources**
  - Data for Intersection Signal Inventory/State
    - Low frequency & with different subtleties & more relational content
  - Current solution:
    - Mongo+ Kafka/ActiveMQ + Postgres

- **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

	Source	Information Type	System
Arterial Data	Pasadena	Intersection signal	Pasadena TMC
	Duarte	Intersection signal	County TMC
	Monrovia	Intersection signal	County TMC
	Arcadia	Intersection signal	Arcadia TMC
	LA County	Intersection signal	County TMC
Freeway Data	Caltrans FW Traffic	Loop sensing	Caltrans ATMS
	Caltrans FW Ramps	Ramp meters	Caltrans ATMS
	Caltrans FW CMS	DMS	Caltrans ATMS
	Caltrans Intersections	Intersection signal	TSMSS
	Caltrans Video	Video	via RIITS
	Caltrans FW Lane closure	Lane status	LCS
	Caltrans incident	Incident	Caltrans ATMS
	210 LCS	Lane status	High speed rail system
Transit Data & Other	RIITS Environmental sensing	Environmental	RIITS
	RIITS Transit	Transit	RIITS
	RIITS Video	Video	RIITS
	Gold line transit	Transit	NextBus
	511 (Out only)	Response plan information	
	Bluetooth traffic	Travel time	County TMC

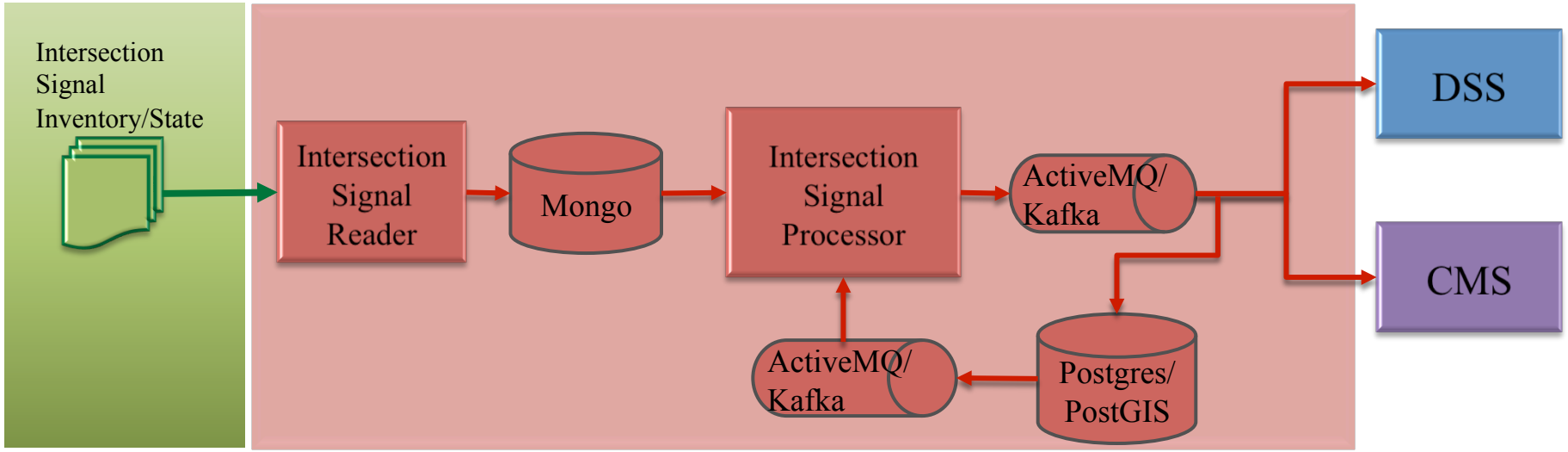
Real-time Data Streams



Data Example: Arterial Sensor Data

DetectorID	Date	Time	State	Speed	Occupancy	Volume	AvgSpeed	AvgOccupancy	AvgVolume
307502	20170100	48541	OPERATIONAL	0.00	0.00	0.00	1.00	50.00	1.00
307502	20170104	34715	OPERATIONAL	15.00	6.00	240.00	21.00	7.00	325.00
307502	20170105	71735	OPERATIONAL	0.00	0.00	0.00	1.00	50.00	1.00
307502	20170109	78063	OPERATIONAL	24.00	10.00	600.00	26.00	9.00	522.00
307502	20170122	82462	OPERATIONAL	19.00	7.00	360.00	21.00	6.00	297.00
307502	20170123	2267	OPERATIONAL	23.00	4.00	240.00	15.00	3.00	137.00
307502	20170123	47839	OPERATIONAL	24.00	8.00	480.00	23.00	13.00	658.00
307502	20170127	28596	OPERATIONAL	0.00	0.00	0.00	1.00	1.00	1.00
307502	20170128	54630	OPERATIONAL	0.00	97.00	30.00	1.00	96.00	52.00
307502	20170129	1510	OPERATIONAL	19.00	24.00	1200.00	19.00	22.00	1043.00

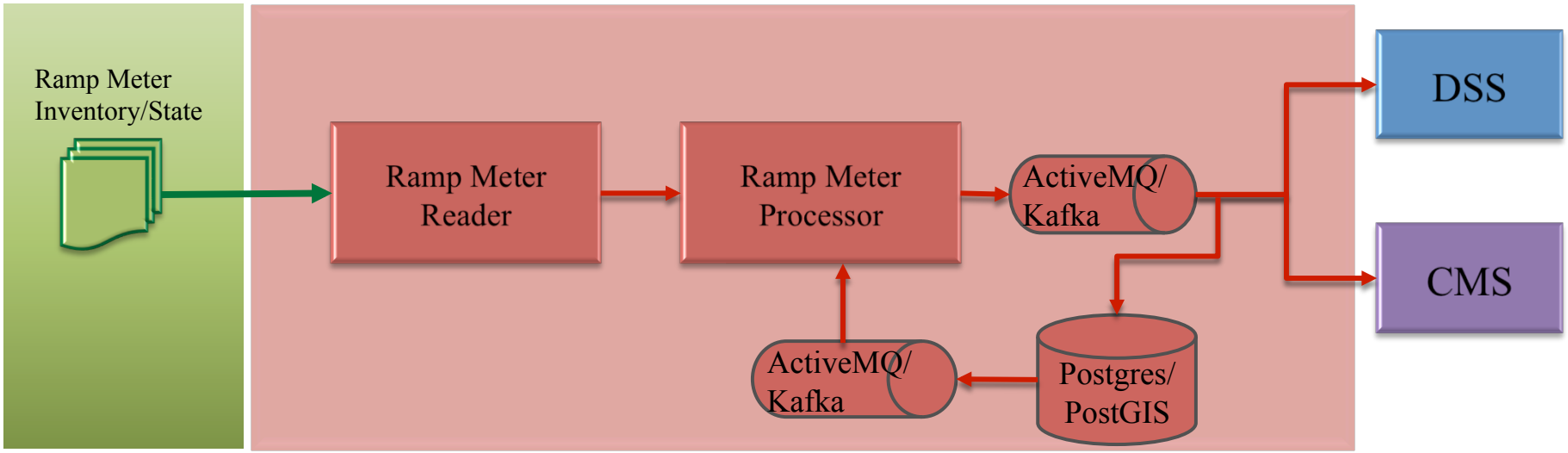
Heterogeneous Sources



Data Example: Intersection Signal Inventory

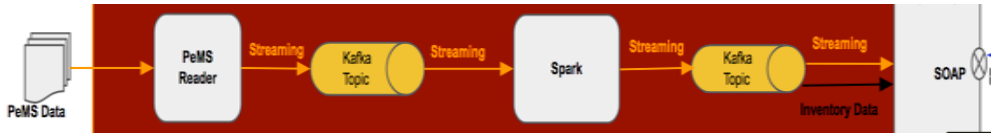
OrgID	IntersectionID	Date	Time	SignalType	Description	MainStreet	CrossStreet	Latitude	Longitude
5:1	3075	20170626	42943	170 LACO IV	Foothill / Second	Foothill Blvd	Second Ave	34.151051	-118.025402
5:1	3076	20170626	42944	170 LACO IV	Foothill / First	Foothill Blvd	First Ave	34.151014	-118.028533
5:1	3077	20170626	42945	2070 D4 1.5L	Foothill / Santa Anita	Foothill Blvd	Santa Anita Ave	34.150945	-118.031604
5:1	3078	20170626	42946	170 LACO IV	Foothill / Baldwin	Foothill Blvd	Baldwin Ave	34.150598	-118.050237
5:1	3079	20170626	42948	170 LACO IV	Colorado / Baldwin	Colorado St	Baldwin Ave	34.148326	-118.050243
5:1	3080	20170626	42887	2070 D4 1.5L	Baldwin / Gate 7	Baldwin Ave	Gate 7	34.142715	-118.051142
5:1	3081	20170626	42888	2070 D4 1.5L	Baldwin / Gate 8	Baldwin	Gate 8	34.138967	-118.053101
5:1	3082	20170626	42889	2070 D4 1.5L	Baldwin / Gate 9	Baldwin	Gate 9	34.136753	-118.054136
5:1	3091	20170626	42890	2070 D4 1.5L	Baldwin / Gate 10	Baldwin	Gate 10	34.134082	-118.054322
5:1	3092	20170626	42891	2070 D4 1.5L	Huntington / Baldwin	Huntington	Baldwin	34.131693	-118.054508
5:1	3093	20170626	42892	2070 D4 1.5L	Huntington / Gate 1	Huntington	Gate 1	34.131808	-118.051978
5:1	3094	20170626	42893	2070 D4 1.5L	Huntington / La Cad...	Huntington	La Cadena	34.13173	-118.049656
5:1	3095	20170626	42894	2070 D4 1.5L	Huntington / Michilli...	Huntington Dr	Michillinda Ave	34.130806	-118.067429
5:1	3096	20170626	42895	2070 D4 1.5L	Huntington / Sunset	Huntington Dr	Sunset Blvd	34.130945	-118.065124

Homogeneous Sources

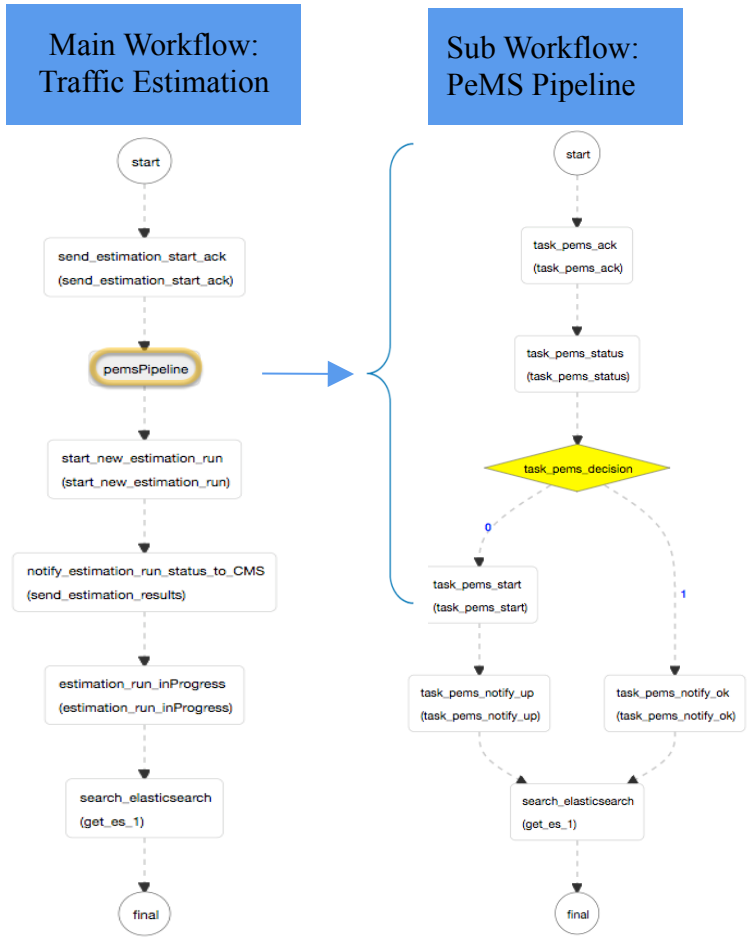


Orchestration – Command Gateway

- **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

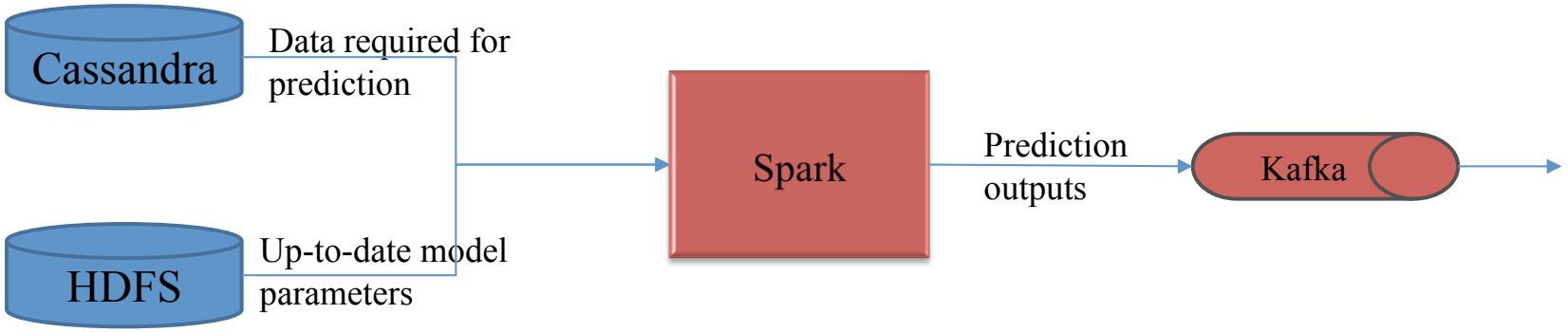


- **External Interface: Independent pipelines with workflow and data flow control**
  - Each data pipeline has no knowledge of the other pipelines or how to communicate with the other pipelines
  - The knowledge of workflow and message routing stays in a central place - Command Gateway



Application Examples

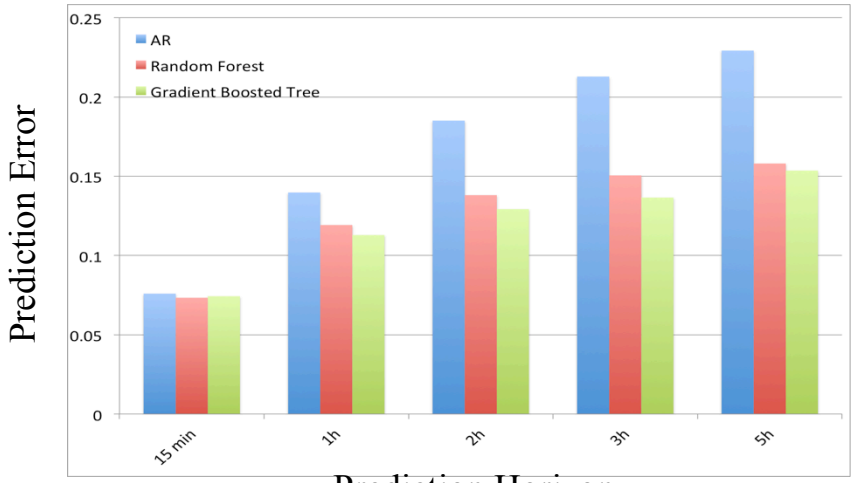
Machine Learning – Flow Prediction



Three models from *Mllib* currently deployed in the Data Hub:

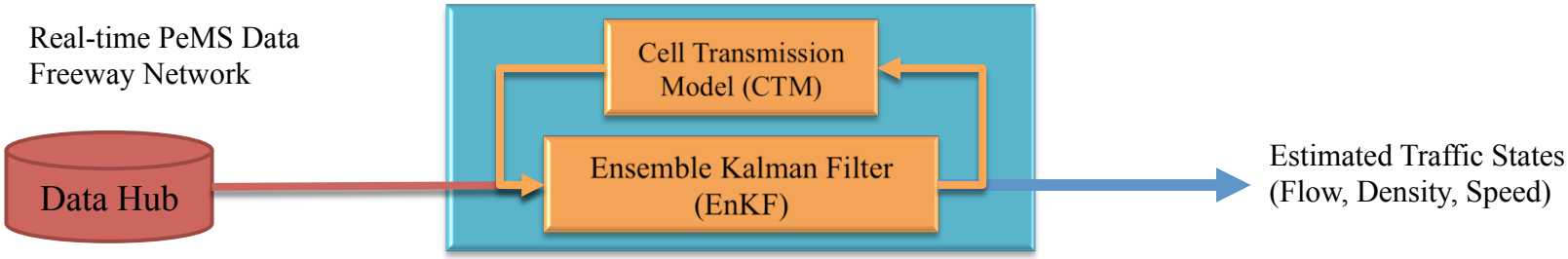
- ❑ Gradient boosted tree **(Best Performance)**
- ❑ Random forest
- ❑ Linear autoregressive model

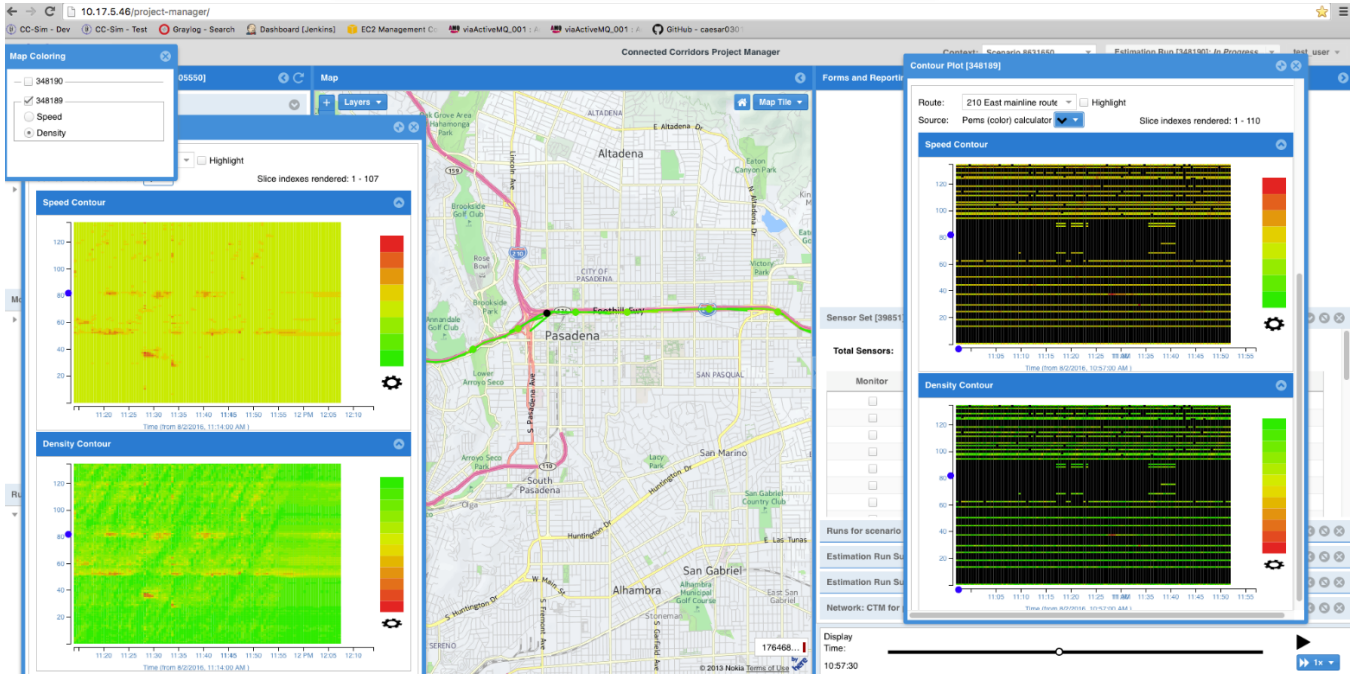
Flow prediction at a mainline sensor



Freeway Traffic Estimation

- Real-time PeMS Data
- Freeway Network

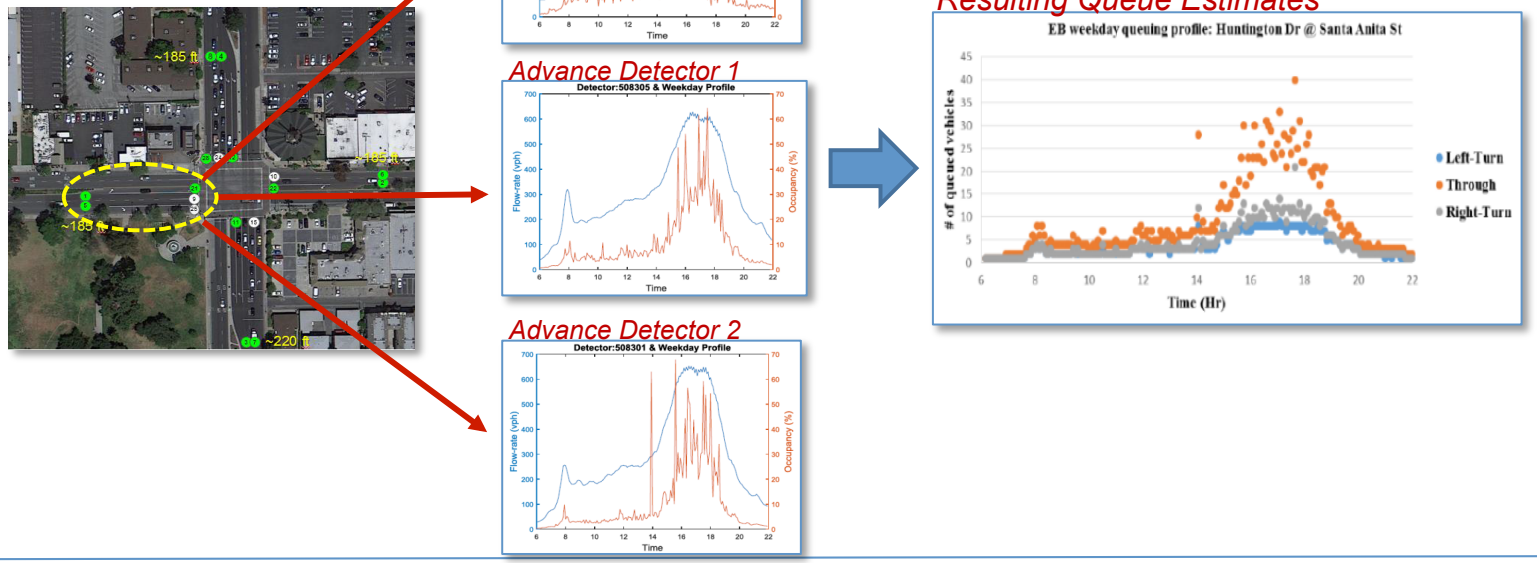




Arterial Queue Estimation

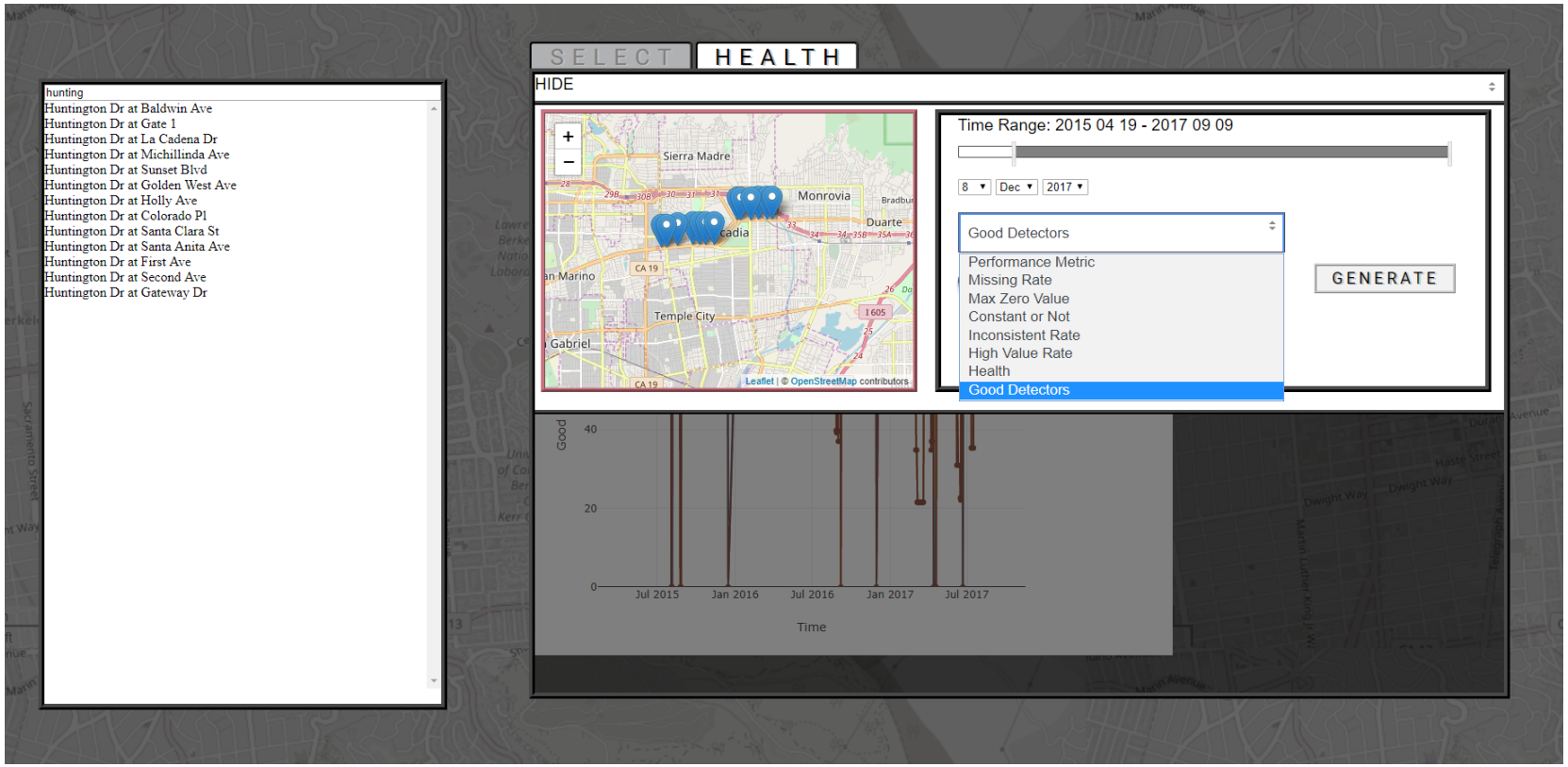
- Arterial Detector Data
- Arterial Signal Phasing Data
- Intersection Road Geometry

Estimated Traffic Queues (Left-Turn, Through, Right-Turn)



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



24

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)

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

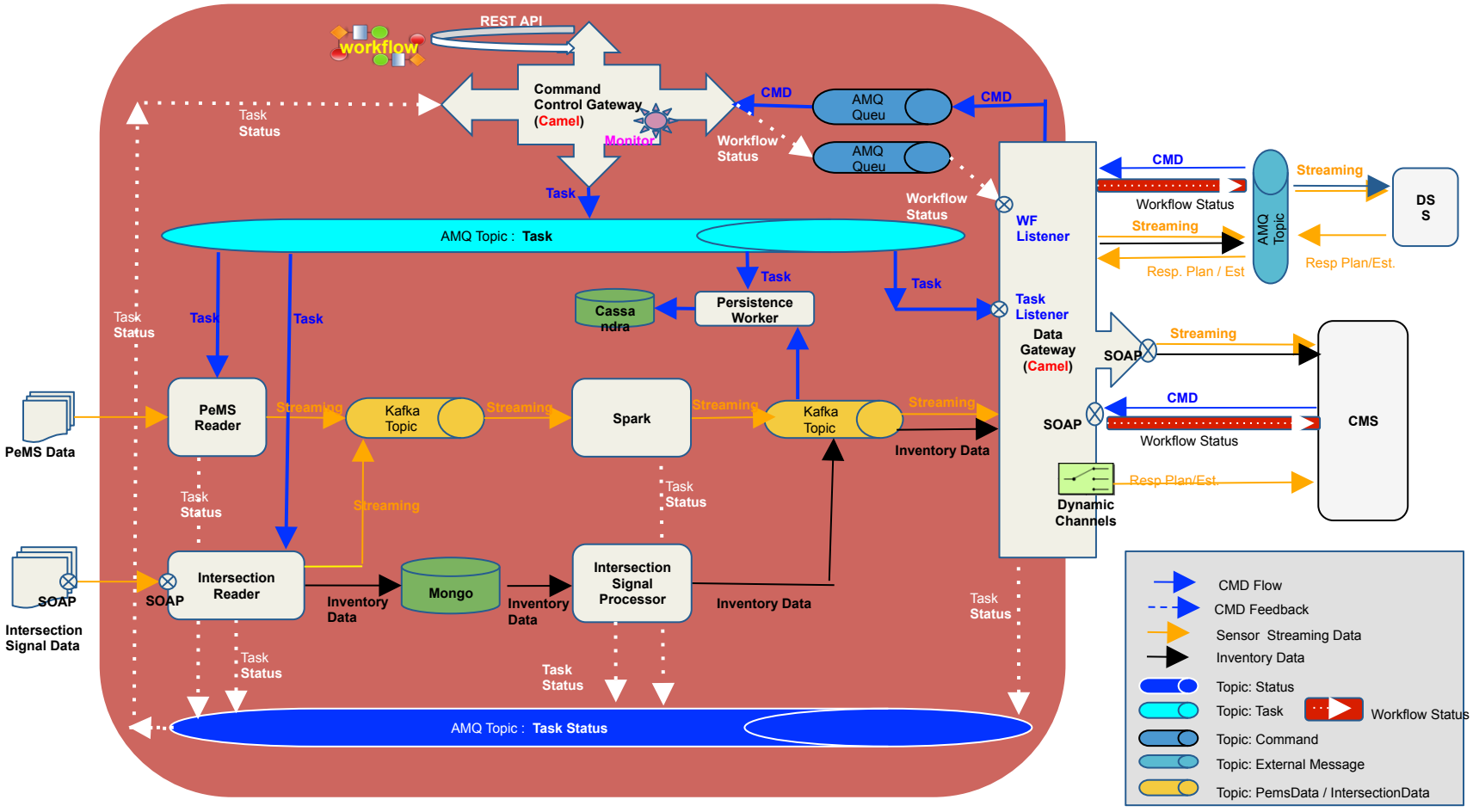
Technology Stack (2/2)

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

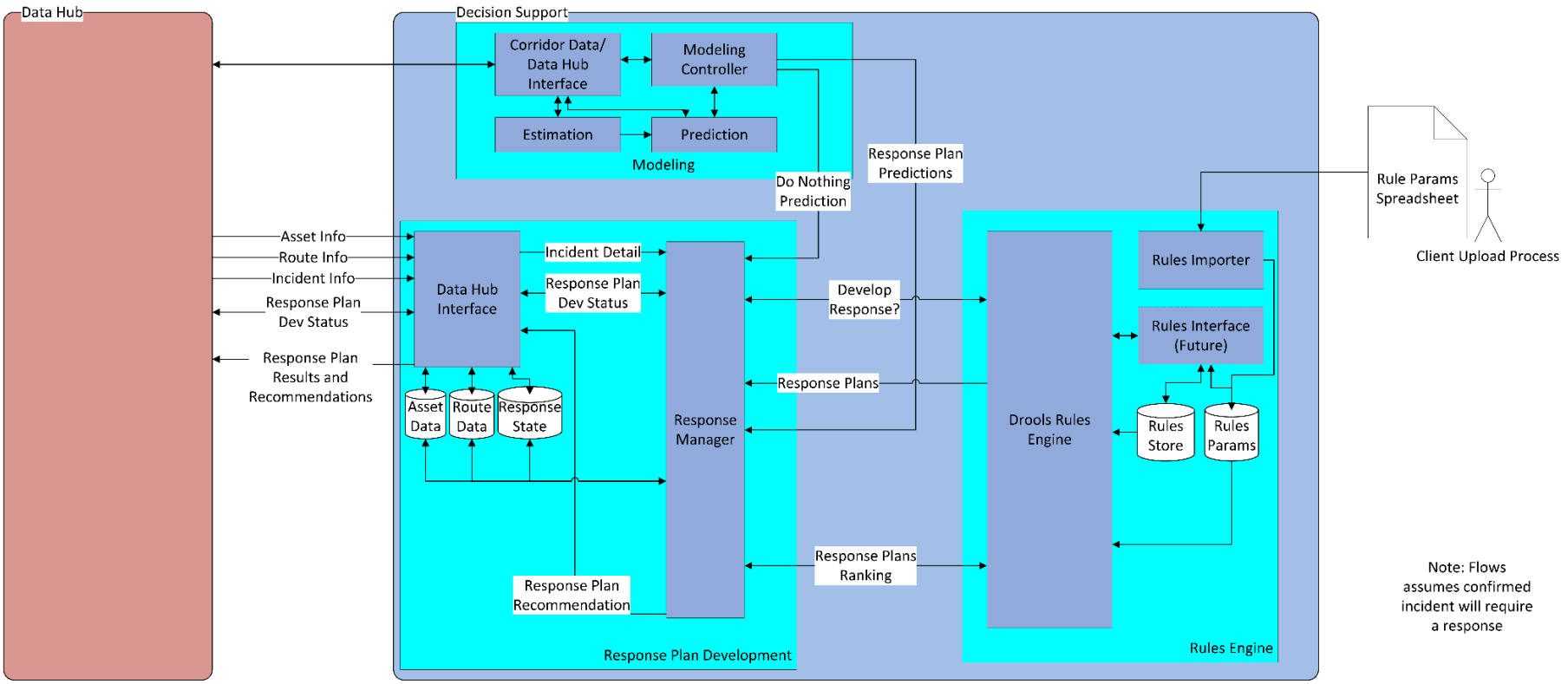
Primary AWS Services

Technology	Purpose	Key uses
EC2	Server processing on demand	Estimation, Prediction, data processing, Persistence workers, Cassandra, MongoDB, other custom workers, messaging, logging
RDS	Postgres w/PostGIS	Modeling data store (models, corridor asset model element information) Data hub relational store (corridor asset post transformation)
S3	Storage	Stateful processing
Security Groups/ VPC/IAM	Cloud/network isolation/identity & access management	Networking/Security/Cloud access
EMR	Hosted Spark	Analytics, data quality, machine learning
Cloud Init, Cloud Formation	Deployment	Instance automation, cloud initialization and maintenance
CloudWatch, CloudTrail	Monitoring	Monitoring
Key Management Service	Key Management	Security, encryption
Code Commit, Code Deploy, Code Build, Code Pipeline	Code repository, Code builds, Configuration management, Code deployment	Dev and Test Environment deployment, continuous integration

Orchestration - Overview

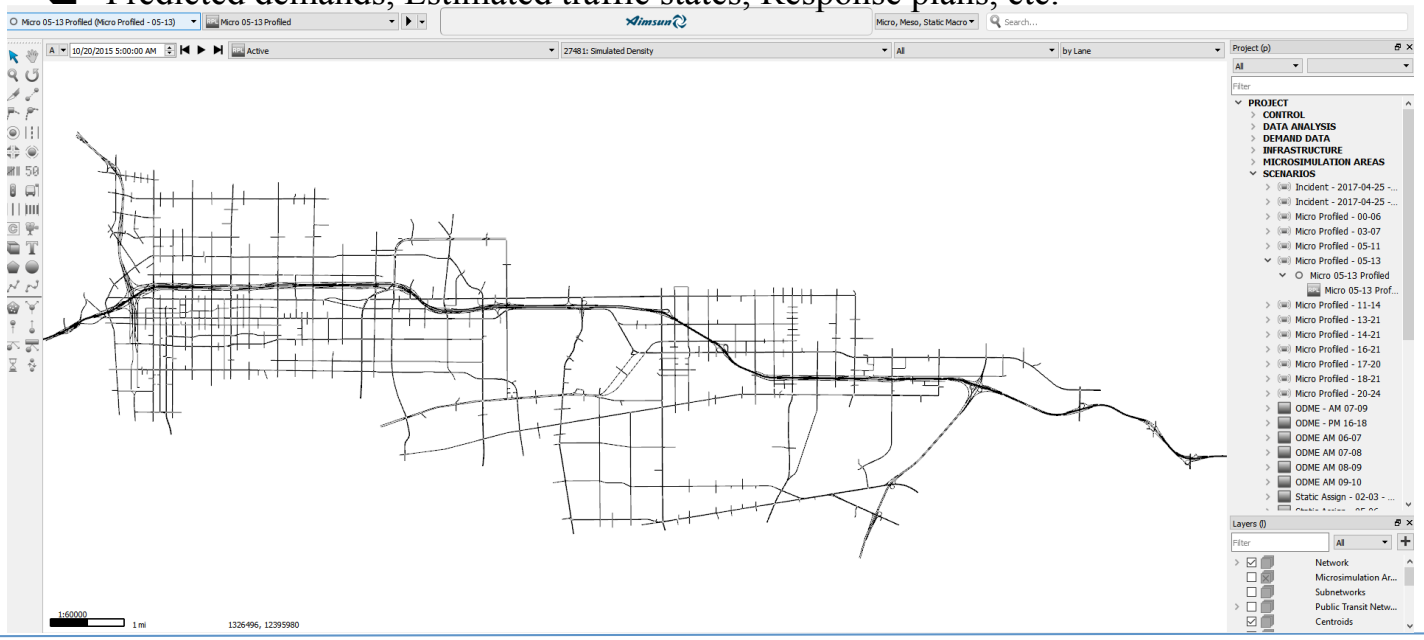


Decision Support System – Design Detail



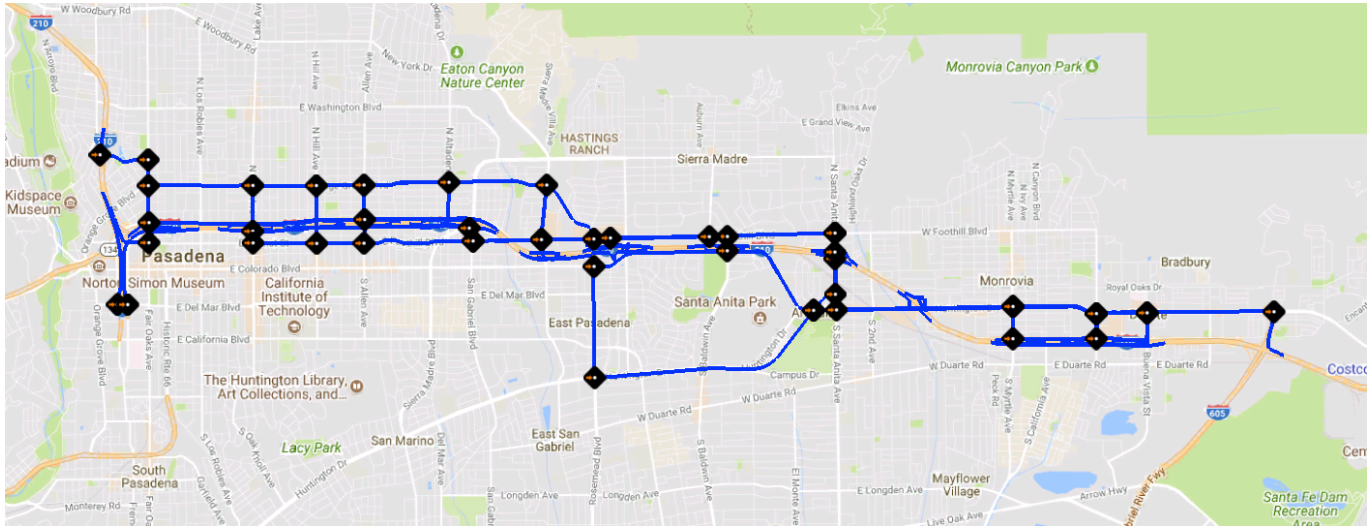
Aimsun Modeling, Calibration, & Prediction

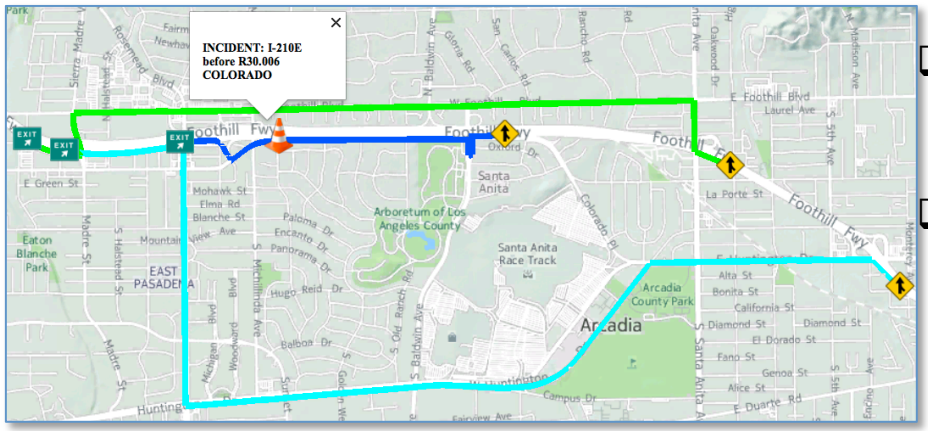
- The Aimsun Model
  - ~1000 lane miles of road, ~5000 traffic detectors, 459 signalized intersections and control plans, 45 freeway ramp meters, Metro gold line and all bus routes
- Data Inputs
  - Current
    - 2008 SCAG data, observed flow counts from the field, signal timing plans, ramp metering 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