Visualizing Autotrader Traffic Using Spark Streaming

Jon Gregg, Cox Automotive

Overview

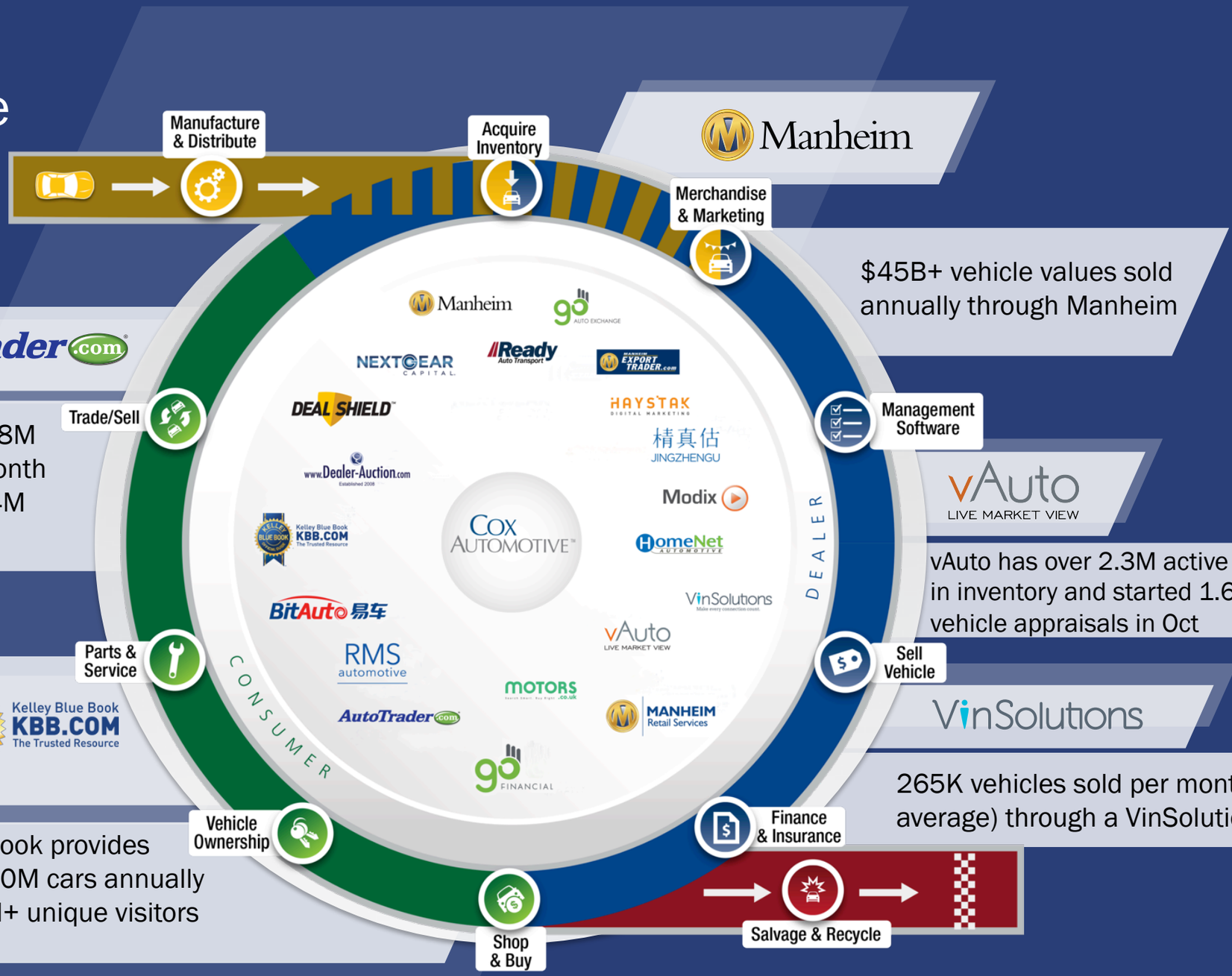
- Cox Automotive and Hadoop
- Spark Streaming application
- Spark roadmap at Cox Automotive

Spark summit 2015

Why we're using Hadoop

Cox Automotive

The Journey



AutoTrader.com has 18M unique visitors each month and lists an average of 4M cars monthly

Kelley Blue Book provides values for 290M cars annually and has 18M+ unique visitors monthly

$45B+ vehicle values sold annually through Manheim

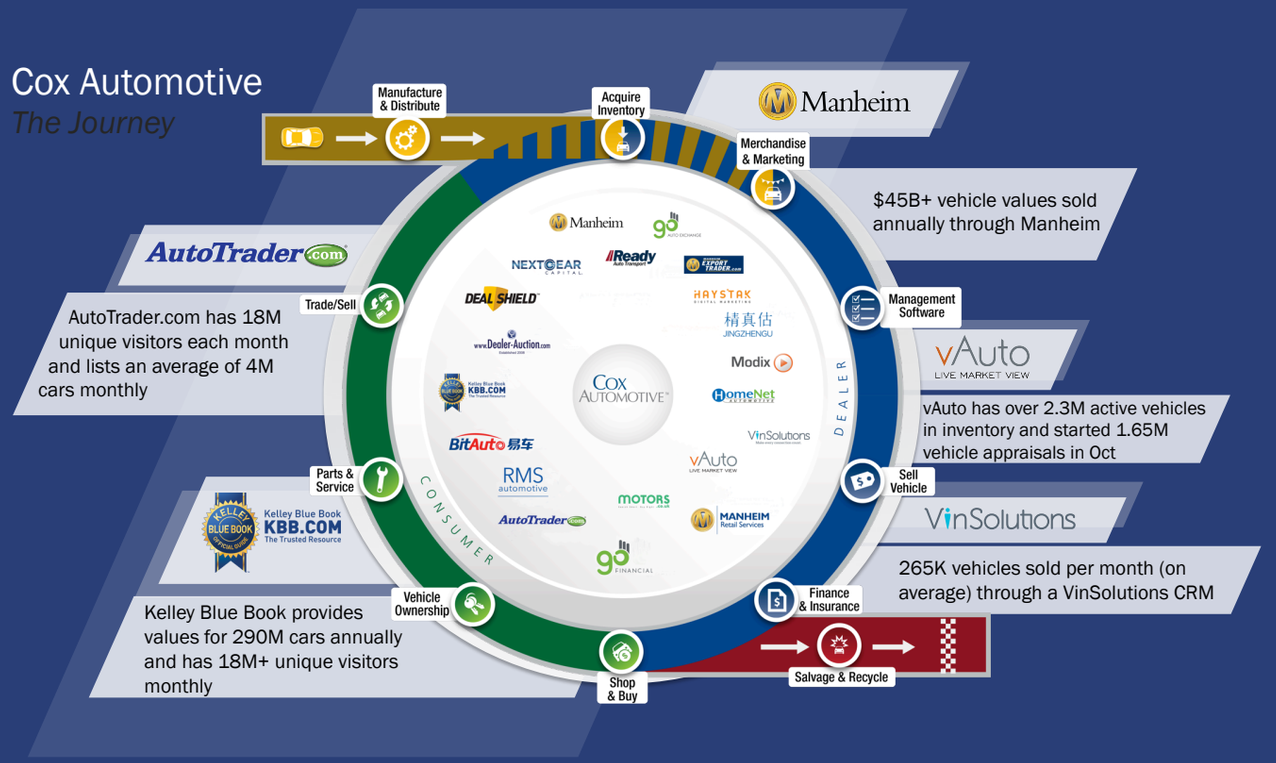
vAuto has over 2.3M active vehicles in inventory and started 1.65M vehicle appraisals in Oct

265K vehicles sold per month (on average) through a VinSolutions CRM

Spark summit 2015

Cox Automotive

- Over 25 companies (and growing)



- Facilitate joining data, analyst collaboration
- Hadoop cluster, dedicated ingest team

Use Hadoop where it makes sense

- Joining data from across several companies
- Large amounts of data (Querying and Reporting)
- Build out business logic so it's shareable

That’s all great…

… but we also have to showcase what Hadoop can do

Autotrader's "Autobowl"

Autobowl: Goal

Find which Big Game car commercial led to the greatest Autotrader traffic increase, as a proxy for influence on consumers?

Spark summit 2015

Two solutions

- Hive on MapReduce
  - Mature, supported product
  - Shows SQL's capabilities on Hadoop
- Spark
  - Started as a POC, no expectations
  - How would it work with YARN, Kerberos?

Autobowl Hourly Data

Make Model	Hour	VDPs	Searches
…			
Kia Sedona	9pm	300	290
Kia Sedona	10pm	310	320
Kia Sorento	3pm	220	240
Kia Sorento	4pm	210	220
Kia Sorento	5pm	350	380
…			

+70%!

Comparison: Hive vs. Spark

Spark summit 2015

Hive vs. Spark: Processing Time

Minutes to Process 1hr of Site Activity Data

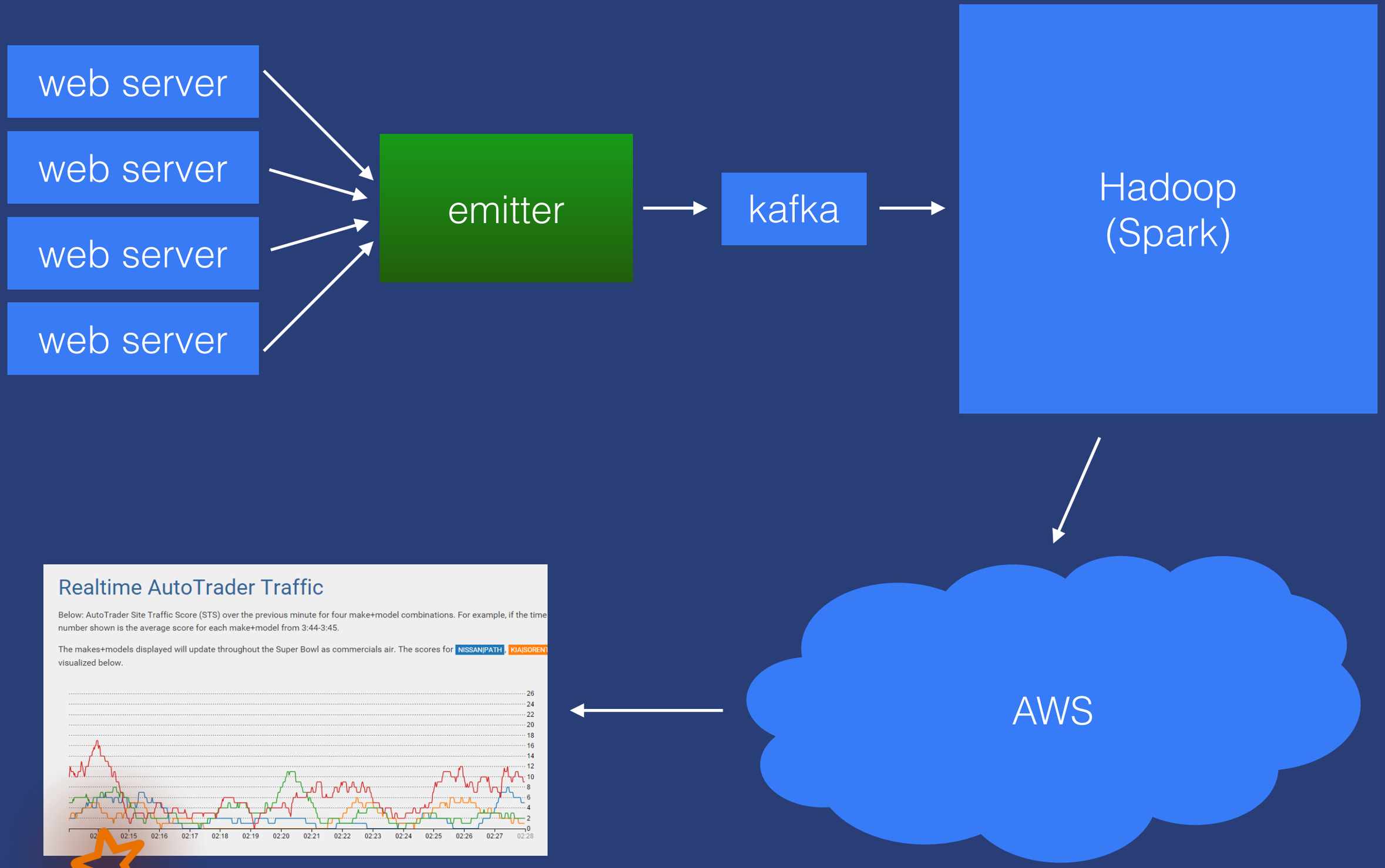
Spark Streaming for Near Realtime Visualization of Traffic

Autobowl Hourly Data

Make Model	Hour	VDPs	Searches
…			
Kia Sedona	9pm	300	290
Kia Sedona	10pm	310	320
Kia Sorento	3pm	220	240
Kia Sorento	4pm	210	220
Kia Sorento	5pm	350	380
…			

How about a visualization using Spark Streaming?

High-level architecture diagram



Spark summit 2015

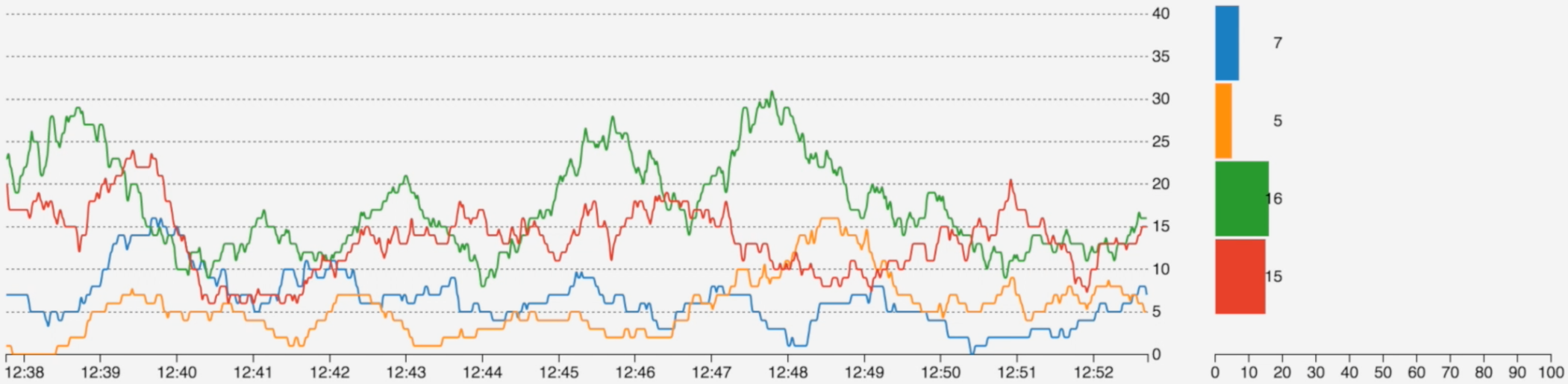
Video

(Screenshot from Video)

Realtime AutoTrader Traffic

Below: AutoTrader Site Traffic Score (STS) over the previous minute for four make+model combinations. For example, if the time shown on the graph is 3:45, then the number shown is the average score for each make+model from 3:44-3:45.

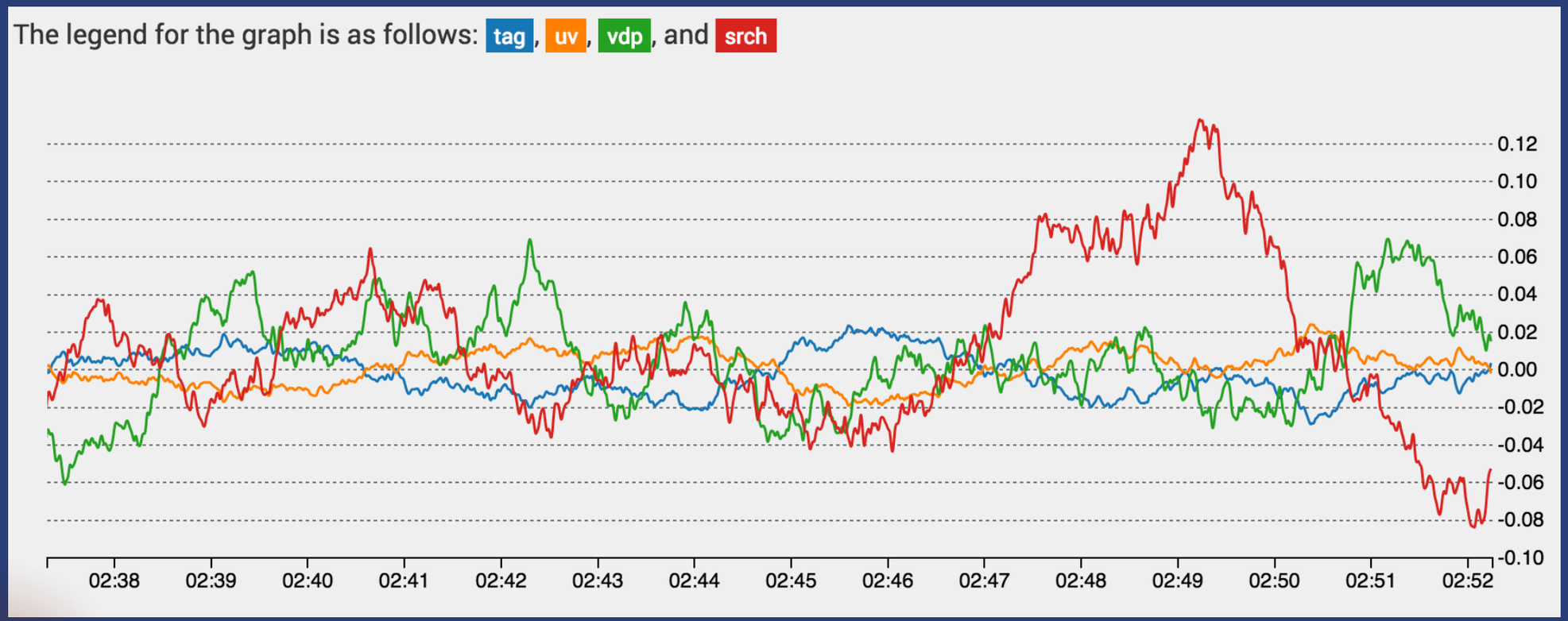
The makes+models displayed will update throughout the Super Bowl as commercials air. The scores for CHEV|COLORADO, KIA|SORENTO, JEEP|JEEPGRAND, and TOYOTA|CAMRY are visualized below.



What's next?

Other Visualization use cases

- Detecting anomalies in Autotrader metrics after a site update



Other Visualization use cases

- Detecting anomalies in Autotrader metrics after a site update
- Executive dashboards
- Visualizations for A/B testing

Gaining BI Adoption

Gaining BI Adoption

- Most BI users use Hive or point-and-click app
- But there's been a shift - Spark is in use by analyst teams within Autotrader, KBB using Python
- Spark used by developers at Autotrader, KBB, Mannheim, NextGear

Python: our primary Spark language

- Speed improvements with Dataframes, Kafka integration
- “Easier sell” than Java/Scala
  - Scripting, visualization+analytics packages
- Onboarding users
  - Central repository with best practices
  - Individual support for BI Spark champions
  - Guidelines on setting Spark parameters