Competing on Analytics: Technology Perspective

IHS Case Study: A Big Data Platform – The Foundation to Advanced Analytics

Michelle Eissele
Senior Director, Data Operations
IHS Automotive delivers a powerful combination of actionable data-driven insights, expertise and knowledge-based solutions to drive competitive advantage for automotive stakeholders.

IHS Automotive receives and processes thousands of different data sources and through systems and experts adds analytics and expertise to deliver insights that drive the Automotive industry.

Products include:
- Product Planning & Strategy: Components, Long-term Planning Scenarios
- Marketing: Audience Targeting, Activation and Optimization
- Market Reporting: Sales Analysis, Sales Forecasts
- Aftermarket: Vehicles in Operation, Aftermarket Planning

In the US we receive household demographic data, vehicle title and registration data, vehicle information, survey data and dealer information.
Our Opportunity

**Margin Opportunity**

- **Descriptive**
  - What happened?
  - Polk Loyalty
- **Diagnostic**
  - Why it happened?
  - IHS Business Analytics
- **Predictive**
  - What will happen?
  - Garage Predictor
  - CO₂ Forecast
- **Prescriptive**
  - How should I respond?
  - Specific consulting engagements

**Includes market reporting products such as PolkInsight**

**Growth Opportunity**
As with most companies, we had built optimized data marts and data warehouses built around answering very specific questions and delivering data to individual products.

In addition, IHS had just acquired R.L. Polk so we had two companies with sub-optimized data warehouses.
Our Vision

- Existing Polk Data
- Existing IHS Data
- New Exploratory Data

Data Analytics Warehouse

- Linking & Integration
- Analytics & Visualization

- Analytics Sandbox
  - Production Analytics
  - Ad Hoc
  - Cyclical Jobs

Outputs to Applications/Customers

Analytics Sandbox

Existing IHS Data

Existing Polk Data

New Exploratory Data

Person
Household
Business

Vehicle

Geography
To create a foundational platform that would serve as the fundamental building block needed to increase our analytic capability and allow our analysts the ability to discover and provide impactful information to our customers

- Key components of the new proposed system:
  - Create a linking strategy to integrate data from both Polk and IHS
  - Anonymize the PII data to increase security and adhere to privacy laws
  - Create ability to bring in new exploratory data sets quickly and easily
  - Provide a platform that is scalable, fast and easy to use
  - Utilize best in class data analysis and visualization tools

- The system will focus on two main goals
  - Creating an analytical sandbox that will allow analysts the ability to easily view and link existing and new data in order to gain valuable insights that will drive value for our customers
  - Creating a production environment that can be leveraged to deliver a variety of analytical solutions to our customers
The Solution

Analytics Toolbox

Level 3 (Aggregated)
- Vehicle Purchase Activity
- VIO Aggregate
- Demo Aggregate
- Statistical Samples

Level 2 (Constructed)
- Current Garage
- VIO Summary
- Historical Garage
- Vehicle History

Level 1 (Raw)
- Vehicle Coding
- RDM
- Vehicle Mileage
- Demo graphics
- Vehicles in Ops
- Owner History
- Loyalty
- PIK Coding
- Vehicle Ref
- CMV
- New Vehicle
- Dealer

Data Linking
- Household
- PIK Key
- Vehicle
- Geography

Process to bring in Exploratory Data

Exploratory Data
- Survey Data

Persisted Analytics Data

Reference Data
- Vehicle Attribute Data
- Ref Data
- Master Dealer Data
- Segment

Core Data Sets
- Vehicles in Ops
- Demos
- New Vehicle
- Loyalty
- Vehicle Event History
- Historical Demos
- Sales Data
- Used Vehicle

Linking Processes by HH Match Key, PIK Key, Vehicle Attribute or Geography
- **Data Organization**
  - Data is built into specific level for ease of access and use
  - Data Engineers build foundational data sets made up of “raw” data facts and construct logic data answers to common questions
  - Analysts build aggregated data sets that are used and shared for analytical jobs and investigations

- **Unparalleled History in One Place**
  - Rich history of vehicle ownership facts
    - Transaction level details cleansed and standardized back to June 1993
    - VIN decoding rules for model years 1961 to present
    - Other persisted reference data such as Demographics and survey data

- **Intelligent Linking that enables the highest level of Security**
  - Householding algorithms allow us to link households over time
  - Protect Identify Key (PIK key) – Completely anonymizes the PII information while preserving the ability to link households
Breadth of Data

**Demographics**
- Income
- Age
- Ethnicity
- Lifestyle Characteristics

**Vehicle Data**
- Owners
- Vehicles
- Locations

**GARAGE COMPOSITION**
- Lifestyle Data
- Demographics
- Ethnicity
- Loyalty Results
- Survey Data

197,123,146 Households
648,113,518 Vehicles
3,201,446,935 Ownership Records

Employment Data / Weather / Fuel Prices / Consumer Sentiment

- Segmentation
  - Brand Loyalty
  - Custom Geo
- New Vehicle
- Sales Data
- Vehicle Attributes
- Dealer Ref Data
History of Data

- **1993**: Home Price = $100K
- **2005**: Moved
  - **New Home Price** = $350K

- **1995**: Used
- **2007**: New Defector
- **2009**: New
- **2011**: New, Loyal Segment Change
- **2013**: New, Loyal Segment Change

- **+1 Child**
  - Liked Car, Didn’t Like Gas Mileage
  - Needed something Bigger
  - Drives great, love car

- **$30K to $150K** Income

- **Weather Data, Incentive Data, Fuel Prices, On-line Shopping, etc**

- **Customer Data**
What worked well

- Data side ownership with technology support – it is all about the data!
- Accelerated tools and platform selection jump started our project effort
- Assigned our best people with strong background in data technologies
- Invested in upfront training and support with Cloudera
- Allowed time to play and discover in the new technology
- Created a structured roll-out and communication plan

What we would do differently next time:

- Involve more users of the system up front in the project
- Set more realistic project timing and scope