

Merchant classification on 30 billion transactions a month, and the API that sold it

Clasificación de comercios sobre 30.000 millones de transacciones mensuales y el producto de datos que la comercializó

Role: Data Scientist · Argus Information & Advisory Services (a Verisk Analytics company), White Plains NY · 2014 – 2016

CONTEXT

Argus aggregated card-transaction data from issuers at enormous scale. The raw merchant descriptors on those transactions were inconsistent free text, which made it impossible to answer the questions media and marketing clients were asking: who shops where, how often, with what share of wallet, and how that changes after a campaign.

Delivering insight to external clients took days of manual work per request.

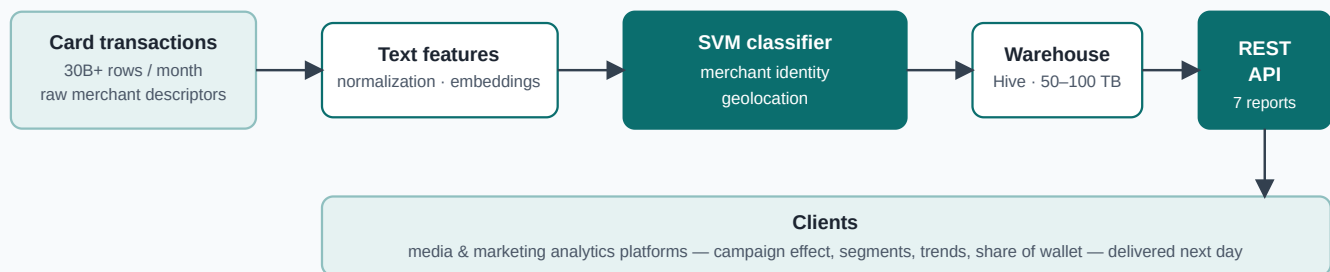
WHAT I BUILT

- A production merchant-classification system: word embeddings over raw transaction descriptors feeding an SVM classifier that resolves merchant identity and geolocation, trained and scored on 30B+ rows per month.
- Consumption-forecasting and analytics models that joined structured card data with unstructured sources.
- A RESTful data product (the Media Measurement API) exposing seven report types across 25 dimensions, so clients could pull trends, segments and campaign effects programmatically.
- Data modeling and processing of 50–100 TB on Hadoop Hive and SQL Server.

OUTCOME

- Adopted for media analytics by Google (DoubleClick 360), Yahoo Finance, Viant and Pandora.
- Client data latency cut from multi-day to **one day post-transaction**.
- Established a reusable, commercial data product rather than one-off analyses; presented it to partners and prospects.
- Foundation for my later work on large-scale text classification and NLP in consumer analytics.

SYSTEM OVERVIEW



Stack: R · Java · Hadoop Hive · Microsoft SQL Server · Tableau

30B+

transactions classified per month

1 day

client data latency, down from multi-day

7 × 25

report types × dimensions exposed by the API

4

marquee media-analytics adopters

STACK

R · Java · SVM · Word embeddings · Hadoop Hive · SQL Server · REST APIs · Tableau

Sanitized: describes methods and my role; no client data, issuer data or proprietary model details.