



Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Scalable Machine Learning Pipeline for Metadata Discovery from eBay Listings

Qing Zhang, Rui Li

eBay

Spark Summit 2016, June 6-8 San Francisco





Table of Contents

Spark for
Metadata
Discovery

- 1 Who We Are
- 2 Metadata Discovery and Challenges
- 3 Spark Solution
- 4 Summary

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary



eBay Structured Data

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

- Metadata discovery and management
- Listing classification
- Catalog and mapping listing to product
- Inventory insights

The screenshot shows the eBay homepage with a search for 'golf clubs'. The top navigation bar includes links for 'Sign in or register', 'Daily Deals', 'Gift Cards', 'Sell', 'Help & Contact', and a 'CELEBRATE YOUR GRAD' banner. The search bar shows 'golf clubs' with a dropdown menu for 'Shop by category'. Below the search bar, there are related search terms: 'mens golf club sets', 'golf club sets', 'golf iron set', 'golf bag', 'hybrid golf clubs', and 'callaway golf clubs'. The left sidebar contains a 'Categories' section with 'Sporting Goods' expanded, showing 'Golf' and 'Golf Clubs & Equipment'. Below this is a 'Club Type' section with a list of categories and their item counts: Complete Club Set (3,245), Driver (91,162), Fairway Wood (59,823), Hybrid, Utility Club (31,485), Single Iron (48,134), Iron Set (45,210), Putter (33,173), and Wedge (54,564). There is also a 'Dexterity' section with 'Left-Handed' (70,670), 'Right-Handed' (275,892), and 'Not Specified' (106,256). The main content area shows 'All Listings', 'Auction', and 'Buy It Now' tabs, with 'All Listings' selected. It displays 'golf clubs' with 453,328 listings and a 'Follow this search' button. A featured section titled 'Top golf brands' lists 'Titleist, TaylorMade, and PING' with 'Free shipping' and a 'SHOP NOW' button. Below this, there is a product listing for 'CALLAWAY STRATA ULTIMATE 18 PIECE MEN C W/BAG RIGHT HAND- 2016' with a price of '\$349.00' and 'Free shipping'.



eBay Structured Data

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

- Metadata discovery and management
- Listing classification
- Catalog and mapping listing to product
- Inventory insights

Best Selling

Trending price range



Nikon D D3300 24.2 MP
Digital SLR Camera - Black



\$367.97 - \$449.74



Canon EOS Rebel T5i / EOS
700D 18.0 MP Digital SLR



\$436.92 - \$534.02



eBay Structured Data

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

- Metadata discovery and management
- Listing classification
- Catalog and mapping listing to product
- Inventory insights



Main photo

[Add from computer](#)

[Add from mobile device](#)

More item specifics

Brand

- Cardini
- Carolee
- Carolyn Pollack
- Carrera y Carrera
- Cartier
- Casio**
- Caterpillar

Gender

Features

Water Resistance Rating

Watch Shape



Metadata Zoom In

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

- Important name-value pairs: *brand* - Dell
- Selling flow item specifics
- Search navigation
- Powers internal applications



Metadata Discovery

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

memory box	weiss	yamaha	fisher-price	generic
modway	duluth trading	mek usa dnm	other	sahara club
gokey	longhorn	outdoor gear	trax	wolverine
sk	spiderman	vintage	mixed	orchard corset

- Highly rely on manual review
- Unfamiliar candidates
- The same candidate appears in multiple categories



Challenges in Metadata Discovery

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Term	Site	Categories
scott james	US	Men's Clothing: Blazers & Sport Coats Men's Clothing: Pants Men's Clothing: Casual Shirts Men's Clothing: Dress Shirts
tiella	US	Chandeliers & Ceiling Fixtures Lighting Parts & Accessories
turf	US	Sports Mem, Cards & Fan Shop: Cards: Football
turf	UK	Collectables: Cigarette/Tea/Gum Cards: Cigarette Cards: Other Cigarette Cards



Data Driven Approach for Brand Discovery

Spark for
Metadata
Discovery

- Utilize seller input item specifics
- Utilize supply demand signals from sellers and buyers
- Training data available from previously reviewed candidates

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Item specifics			
Condition:	New: A brand-new, unused, unopened, undamaged item in its original packaging (where packaging is ... Read more	Brand:	Nikon
Optical Zoom:	3x	Model:	D5500
Battery Type:	Lithium-ion	Series:	Nikon D
Connectivity:	USB	MPN:	1546
Color:	Black	Type:	Digital SLR
Bundled Items:	Case or Bag, Flash, Lens, Lens Cleaning Kit, Lens Filter, Memory Card, Memory Reader, Strap (Neck or Wrist), Tripod	Megapixels:	24.2 MP
Manufacturer Warranty:	No	UPC:	Does not apply



Supervised Machine Learning Approach

Spark for
Metadata
Discovery

Who We Are

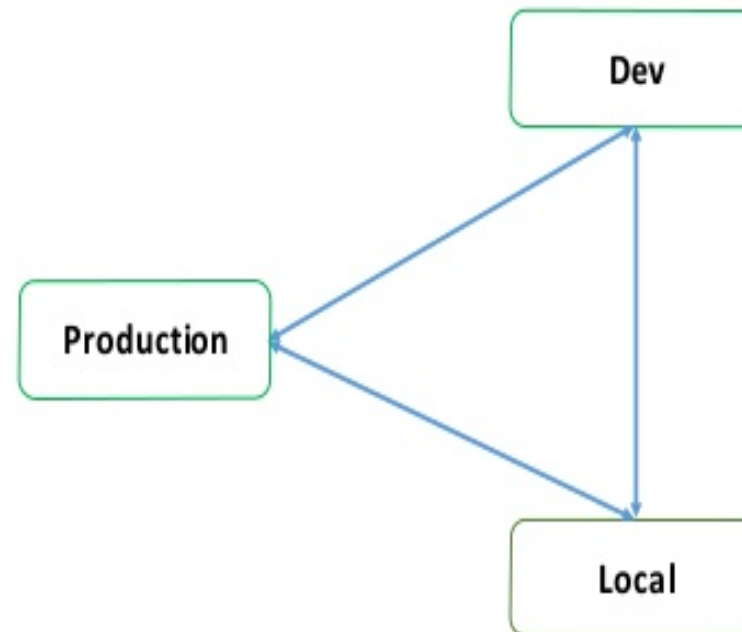
Metadata
Discovery and
Challenges

Spark Solution

Summary

- Data : 35,000 previously human reviewed metadata candidates
- Feature : supply and demand signals
- Prototypes with Python Scikit
- Logistic regression, gradient boosting trees, random forest etc
- Random forest F1 0.878

- In the past, train offline and implement prediction component on production
- File transferring and configurations are time-consuming



MLlib

GraphX

Spark

 **hadoop**

- Spark provides powerful data processing APIs
- MLlib is a comprehensive machine learning package powered by Spark
- Regression, classification, clustering, dimensionality reduction etc
- Efficient development with local model and flexible file access



The Machine Learning System

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Feature

Training

Prediction

MLlib

Spark



Model Training with MLlib

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

```
val pipeline = new Pipeline()
    .setStages(Array(labelIndexer, featureIndexer,
                      rf, labelConverter))

val evaluator = new MulticlassClassificationEvaluator()
    .setLabelCol("indexedLabel")
    .setPredictionCol("prediction")

val cv = new CrossValidator()
    .setEstimator(pipeline)
    .setEvaluator(evaluator)
    .setEstimatorParamMaps(paramGrid)
    .setNumFolds(5)
```


Model	F1
Python Scikit prototype	0.878
MLlib local	0.865
MLlib Hadoop (200 executors, production)	0.862
MLlib Hadoop (400 executors)	0.861
MLlib Hadoop (50 executors)	0.857
MLlib Hadoop (2 executors)	0.862

- The performance variations among implementations are acceptable



Speed

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Stage	Data Size	Time
Feature Generation	1.73 Billion	6 min
Train	33,000	8 min
Prediction Input	650,000	4 min

- Capable of running the job daily
- Speed up the metadata discovery process, from months to days



Newly Discovered Brand

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Brand	Probability
Milkies	0.84
BEABA	0.85
OXO	0.83
Lorex	0.87
Plan Toys	0.85
Safety 1st	0.82
Blabla	0.81
Combi	0.88
Graco	0.88
TotsBots	0.85
Realtree	0.85



Summary

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

- Spark enables fast iterations of ML application development
- MLlib is comprehensive, and well integrated with Spark framework
- Dev and test locally, straightforward production deployment
- Compact code : 600 lines
- Need better understanding of the ML algorithm implementations in MLlib



Acknowledgement

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Thejas Durgam

Anu Mandalam

Meital Tahar Zahav & eBay SDO Team

Jean-David Ruvini



Thank You!

Spark for
Metadata
Discovery

Who We Are

Metadata
Discovery and
Challenges

Spark Solution

Summary

Qing Zhang, qzhang12@ebay.com

Rui Li, ruili1@ebay.com