Tailored for Spark

Apache Big Data North America 2016
Petr Igrevskiy
eBay
How to tailor Spark for maximum impact

1. Optimal infrastructure

2. Customized user experience
Outline

1. eBay, Analytics, Hadoop, and Spark
2. Spark Opportunities at eBay
3. QA
BACKGROUND
<table>
<thead>
<tr>
<th>EBAY INC.</th>
<th>EBAY MARKETPLACE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>162 MILLION BUYERS</strong></td>
<td>~800 MILLION LISTINGS</td>
</tr>
<tr>
<td>Number of active eBay Inc. buyers worldwide **</td>
<td>Approximate number of eBay Marketplace listings</td>
</tr>
<tr>
<td><strong>$21.9 BILLION</strong></td>
<td><strong>63% SHIP FOR FREE</strong></td>
</tr>
<tr>
<td>Amount of eBay Inc. GMV – the total value of successfully closed transactions – in Q4 2015 **</td>
<td>Percentage of transactions from eBay that ship for free (reflects U.S., UK and DE)</td>
</tr>
<tr>
<td><strong>$2.3 BILLION</strong></td>
<td><strong>~85% FIXED PRICE</strong></td>
</tr>
<tr>
<td>Amount of eBay Inc. revenue for Q4 2015</td>
<td>Percentage eBay Inc. GMV that is fixed price</td>
</tr>
<tr>
<td><strong>57% INTERNATIONAL</strong></td>
<td><strong>43% MOBILE GMV</strong></td>
</tr>
<tr>
<td>Percentage of eBay Inc. revenue that is international</td>
<td>Percentage of eBay Inc. GMV that was transacted via mobile in Q4 2015</td>
</tr>
<tr>
<td><strong>43% MOBILE GMV</strong></td>
<td><strong>80% NEW ITEMS</strong></td>
</tr>
<tr>
<td><strong>80% NEW ITEMS</strong></td>
<td>Percentage of items on eBay that are sold as new</td>
</tr>
</tbody>
</table>

Q4 2015
Hadoop at eBay

1. Search Index
2. Log Management
3. Operation Metric Management
4. Analytics
Hadoop Hardware

Multiple Generations

12-18 Cores

72-128GB RAM

24-72TB Storage

Provisioned by cabinet
Spark at eBay

• Uses
  – Spark 1.4 to Spark 1.6

• Methods
  – Yarn

• Current utilization
  – 20% analytic clusters

• Use Cases
  – Purchase Suggestions
  – Marketing Optimization
  – Customer Interests, Consistency, and Similarity
  – Kylin Cube Building
Spark Challenges

- Capacity Management and Efficiency
  - Map Reduce => Yarn
  - Job Sizing
- Support
  - Missing vendor support
  - Missing expertise
- Deployment
  - Library conflicts
  - Configuration challenges
  - Distribution sprawl
- Integration
  - Configuration
TAILORING SPARK

Simple things should be simple. Complex things should be possible.

Alan Kay
We can

• Copy
• Test
• Run
Opportunities for Spark

• Flexibility
• Usability
• Simplicity
• Speed
• Transparency
On YARN

- Security
- Multitenancy
- Reliability
- Experience
- Performance
Does it fit?

- Compute
- Storage
- Network
- Provisioning

- Shared Compute resources
- Independently scalable storage
- Flat Network

eBay - Tailored for Spark
Can we make it feel better?

- Standard ADLC
- Test to your level of comfort
- Single click deployment
- Watch every step
- Certify your job
- Let it run
- Did you say UI?

Diagram:

- Register
- Development
- Test
- Packaging
- Certification
- Runtime

- Repos
- CI
- Metadata DB
- Provisioning
- Runtime farm
- Orchestrator
Q/A