Spark and Hadoop at Yahoo: Brought to you by YARN

Andy Feng
Yahoo! Hadoop
(afeng@yahoo-inc.com)
Big-Data in Yahoo!

Users

Raw User Event Data

Categorized User Activity

Analytics

Targeting

Personalization

Inferred User Knowledge
Hadoop + Spark: Empowered by YARN

<table>
<thead>
<tr>
<th>Hadoop: Batch Processing</th>
<th>Spark: Iterative Processing</th>
<th>Storm: Stream Processing</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>YARN: Resource Manager on Grid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage: HDFS, HBase etc</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

30k+ Yahoo! production nodes on YARN since Q1 2013
Shark Pilot: Advertising Data Analytics

- **Business questions**
  - Are two sets of audience cohorts similar to each other?
  - What audience segment is most likely to be interested in this ad campaign?
  - In what way was the new front page rollout different than the previous front page as far as audience engagement goes?
  - What are the right metrics to define user engagement?

- **Shark pilot**
  - 50 nodes, each w/ 96GB RAM
    - Currently loaded w/ 3.2 TB sample data in memory
  - Homegrown BI tools for ad-hoc queries
    - Using *Shark Server* (contributed to community by Yahoo!)
Shark Perf: TCP-H Benchmark

Average Seconds

- Hive 0.11 (100GB text)
- Hive 0.10 (100 GB text)
- Hive 0.10 (100GB RCFile) with matched nMaps/
- Hive 0.11 (100GB equivalent ORCFile) with matched
- Shark 0.7 with cache enabled
- Shark 0.7 without cache enabled
Spark Pilot: Model Training Pipeline

- A DAG of M/R jobs in Hadoop Streaming
  - Feature extraction
  - Train models
  - Score and analyze models

- Initial Spark prototype
  - 3x speedup on feature extraction

- Production launch
  - Apply Spark against complete pipeline
  - Spark on 80 node cluster
    - Thanks to the enhanced UI and metrics in Spark 0.8
Use Case: Ad Targeting

User → Yahoo! Property → User Profile Store → User Profile Builder

User Profile Builder: M/R and Storm

Model Training: Spark

Model
Use Case: Content Recommendation w/ Collaborative Filtering

Input

CF Learning

Ranking

Output

User & story Features

Activity Events

Long-term Learner
Spark

Long-term Learner
Spark

Long-term Model

Short-term Model

Content Ranker

Top-K stories for each user
Spark-YARN: Deployment Simplified

run spark.deploy.yarn.Client --jar ... --class ... --args ...

--queue ... --num-workers ... --worker-memory ...

Spark-YARN (contributed by Yahoo!) is being adopted by community (ex. Taobao) for production use. You should try it on your Hadoop cluster.
Acknowledgement

▪ AMPLab team
  › Outstanding collaboration: Ion, Matei, Reynold, Patrick, Matt, …

▪ Yahoo! Hadoop team
  › Thomas, Bobby, Paul, Rajiv, Mithun, …

▪ Yahoo! Lab.
  › Mridul, Nathan, …

▪ Yahoo! data analytics
  › Supreeth, Ram, Tim, …

▪ Yahoo! spark users
  › Gavin, Jay, Hirakendu, …
We Are Hiring!

http://careers.yahoo.com/