ACCELERATING APACHE SPARK-BASED ANALYTICS ON INTEL ARCHITECTURE

Michael A. Greene
Vice President, Software and Services Group
Intel Corporation

Follow me on Twitter: **@greene1of5**

INTEL SOFTWARE

When everything computes and connects, Intel software relates.

Connected Devices

Connective Fabric

Cloud and Data Center

Ecosystem Enabling

OPEN INTERCONNECT CONSORTIUM™

intel Intel Developer Zone

intel Services

python™

Java

OpenJDK

openstack™ CLOUD SOFTWARE

hadoop

Spark

OS Enabling

1101011010100101010101011100011010100110110010101010110001110010101001010100

Other names and brands may be claimed as the property of others

intel
Software

In a world where everything computes
and connects, Intel® Software relates

“HARDWARE IS THE MEDIUM THROUGH WHICH AMAZING SOFTWARE EXPERIENCES ARE REALIZED”

BETTER TOGETHER

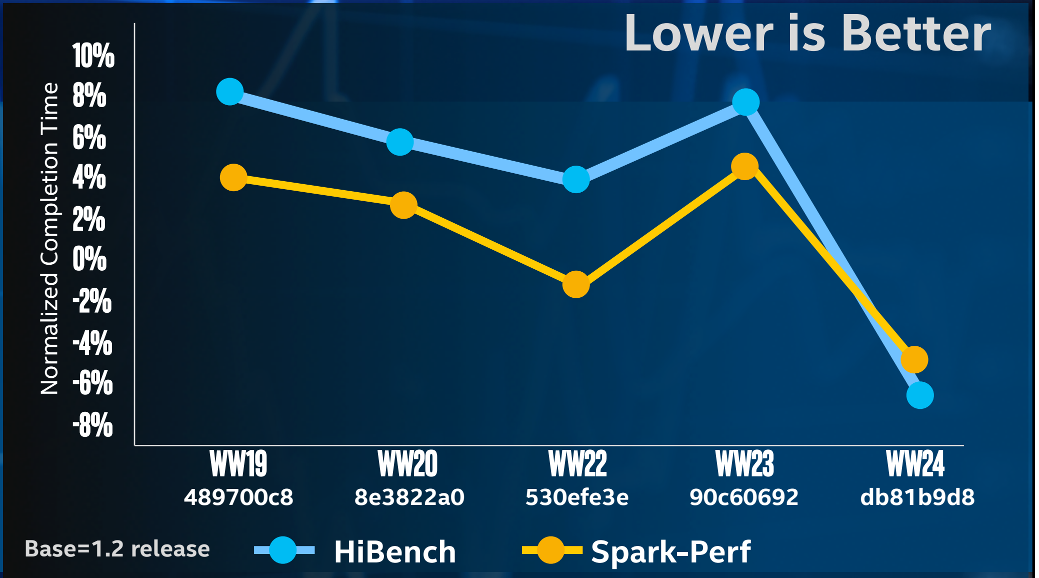
INTEL

+

COMMUNITY

PERFORMANCE PORTAL FOR APACHE SPARK

Weekly Performance Assessment of Spark Upstream



OVERALL PERFORMANCE

http://01org.github.io/sparkscore/plaf1.html

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. Configuration: Intel® Xeon® E5-2697 v2 with 128 GB RAM running JAVA 1.8.0_25, CDH 5.3.2 and the latest version of Spark upstream for stated timeframes.

BETTER TOGETHER

INTEL

+

CUSTOMERS

YOUKU 优酷

Data Processing Time Reduced

94%

Computational Time Reduced

92%

Spark

intel inside Xeon

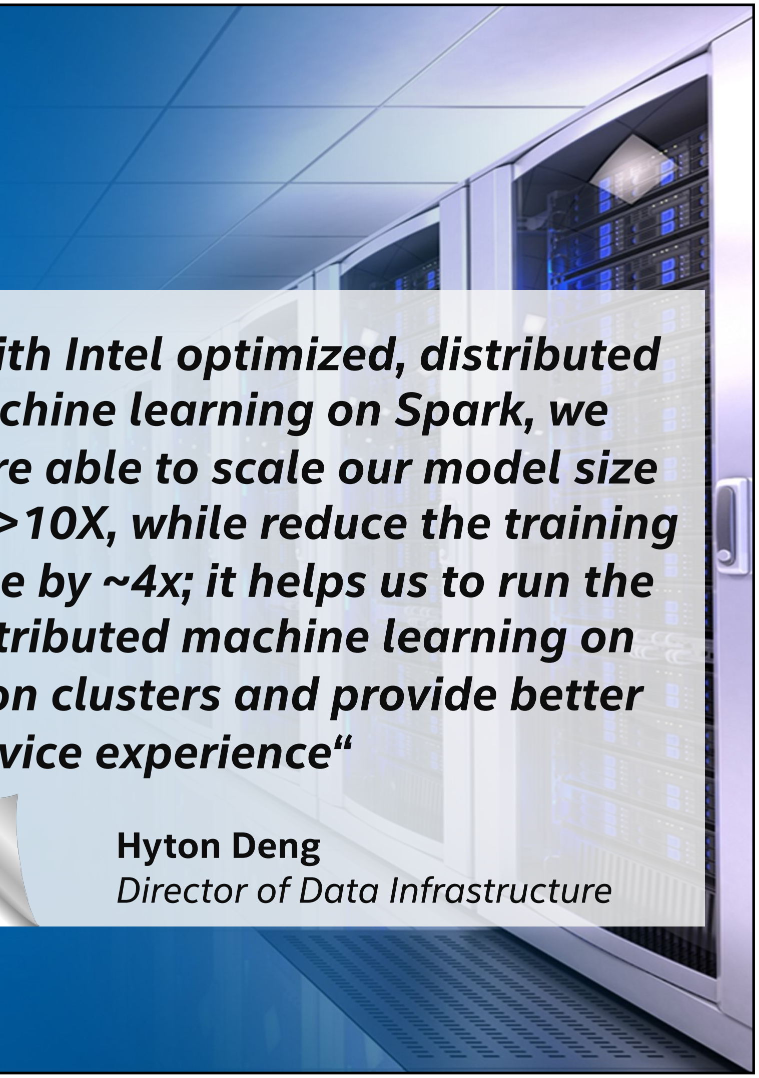
Intel Math Kernel Library

Other names and brands may be claimed as the property of others.
All performance tests were performed and are being reported by Youku Tudou. Please contact Youku Tudou for more information on any performance test reported here.

Spark

intel inside™ Xeon®

Tencent 腾讯



10X

MODEL SIZE

4X

SPEEDUP

> ***“With Intel optimized, distributed machine learning on Spark, we were able to scale our model size by >10X, while reduce the training time by ~4x; it helps us to run the distributed machine learning on Xeon clusters and provide better service experience“***

Hyton Deng
Director of Data Infrastructure

Other names and brands may be claimed as the property of others.
All performance tests were performed and are being reported by Tencent. Please contact Tencent for more information on any performance test reported here.

BETTER TOGETHER

INTEL

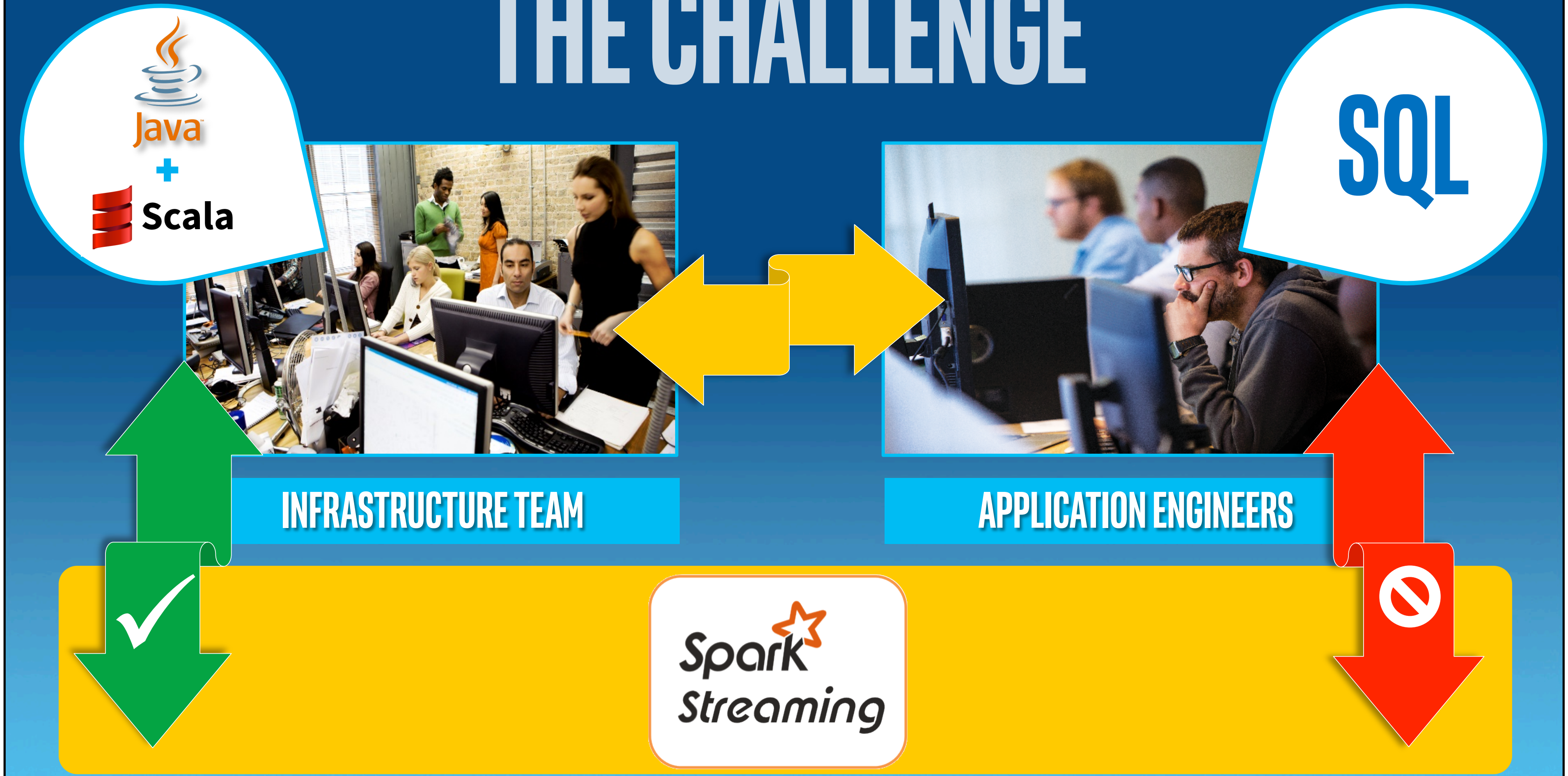
+

ECOSYSTEM

> *“It sometimes took weeks to write new business applications. It was a cumbersome process involving two different teams to analyze real time streaming data.”*

Xiaohui, Liao
Senior Development Manager

THE CHALLENGE



Java + Scala

SQL

INFRASTRUCTURE TEAM

APPLICATION ENGINEERS

Spark Streaming

✓

🚫

STRUCTURED QUERIES

ANNOUNCING:

STREAMING SQL

FOR APACHE SPARK

STREAM PROCESSING

Other names and brands may be claimed as the property of others

Spark Streaming

```
val weblog= KafkaUtils.createStream(…)
 .map(_._2)
 .flapMap(_.split(“|”)).map(_._2, _)

val category=
  sc.textFile(…).flatMap(_.split(“|”))

val result
  = weblog.transform(r =>
r.join(category))
   .map(_._2._2, 1L))
   .reduceByKey(_ + _)

result.print()
```

Streaming SQL

```
CREATE TABLE weblog ( sourcep_ip STRING,
cookie_id INT, item_id INT …) USING
stream.source.KafkaSource OPTIONS( zkQuorum
“localhost:2181”, topic “weblog:1”…)

CREATE TABLE category (item_id INT,
category STRING);

SELECT category, COUNT(*) FROM weblog JOIN
category ON category.item_id =
weblog.item_id GROUP BY category.category;
```

48% of developers use SQL

*Source: StackOverflow.com 2015 survey

THE SOLUTION

Java + Scala

INFRASTRUCTURE TEAM

SQL

APPLICATION ENGINEERS

✓ Spark Streaming ✓

Other names and brands may be claimed as the property of others

“With Streaming SQL for Apache Spark, our Analysts were able to develop business applications in a matter of hours – a huge boost to productivity.”

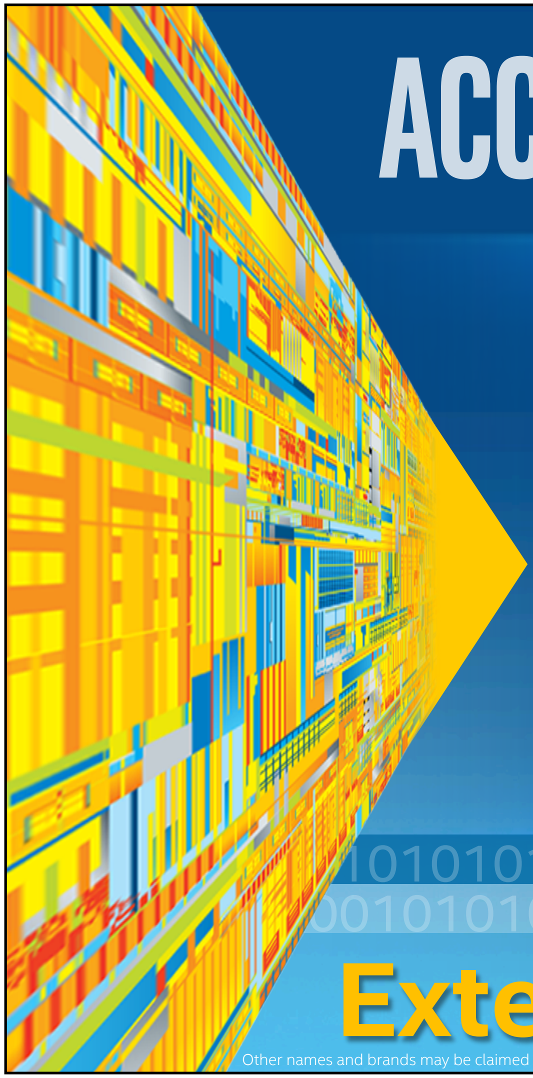
Xiaohui, Liao
Senior Development Manager

JD.COM 京东

JOIN US AND CONTRIBUTE!

https://github.com/Intel-bigdata/spark-streamingsql

ACCELERATING ANALYTICS



SparkR
Streaming SQL
Performance Portal
Project Tungsten

Extend • Stabilize • Optimize

Other names and brands may be claimed as the property of others

BETTER TOGETHER

INTEL + COMMUNITY + CUSTOMERS + ECOSYSTEM

Find out more:
software.intel.com/bigdata

DELIVERING ON THE PROMISE OF BIG DATA

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

experience
what's inside™

Follow me on Twitter: **@greene1of5**