CASE STUDY
Intel® Xeon® Processor E5-2600 Product Family
Big Data Analytics
Communications/Media

Big data enables a personalized video-sharing experience

Youku Tudou Inc. improves its video-sharing application with a big data computing platform based on Intel® Xeon® processor E5-2600 product family, enhancing the user experience and enriching the social media presence of its customers

YOUKU 优酷 土豆 tudou.com

Youku Tudou is China's leading video-sharing website, with average daily views of about 200 million. In keeping with its commitment to provide fast yet reliable service, Youku's average daily visits allow 400 to 500 million responses with an average response time of three to four milliseconds. Based on the new-generation Intel® Xeon® processor E5-2600 product family, it provides improved video-sharing services and a new, more efficient, and targeted user experience for customers.

CHALLENGES

- **Address deficiencies in the distributed computing framework (MapReduce*) when dealing with iterative computation.** Improve iterative computation under the MapReduce distributed computing framework by meeting the requirements of big data processing under the traditional Hadoop* network.
- **Improve performance and efficiency for big data solution operations.** Enhance the video-sharing application's processing performance and efficiency through improved technology solutions for big data applications.

SOLUTION

- **Deploy Intel Xeon processor E5-2600-based Spark/Shark big data computing platform.** Integrate the Spark/Shark big data computing platform into the video-sharing application's own Hadoop cluster to process large amounts of iterative computation through a server platform based on the Intel Xeon processor E5-2600.

IMPACT

- **Improved the video sharing platform's system computation and processing performance.** As one of the first companies commercially applying the Spark/Shark framework in the field, Youku has greatly improved its processing speed and efficiency while enhancing its ability to carry out big data analytics, providing its users with a better video-sharing experiences and enhanced services.

> "With the help of the Spark/Shark* big data computing platform, based on the server platform running on the new-generation Intel® Xeon® processor E5-2600 product family, Youku has been able to improve the performance of big data analytics and computation of its video-sharing application. It has also decreased computation delays to support iterative computation, which requires higher efficiency and machine learning ability. Through this improvement, we were able to enhance efficiency and provide our customers with better video-sharing experiences."
>
> *Lu Xueyu*
> *Technical Director, Big Data Team*
> *Youku Tudou Inc.*

More and more companies have been carrying out detailed analyses of their users' behavior with the help of big data to provide targeted services and improved user experiences. Youku generates huge amounts of data from its large user base. To date, it has collected an estimated 200 million videos, requiring it to process around 200 terabytes of data and 10 trillion records daily.

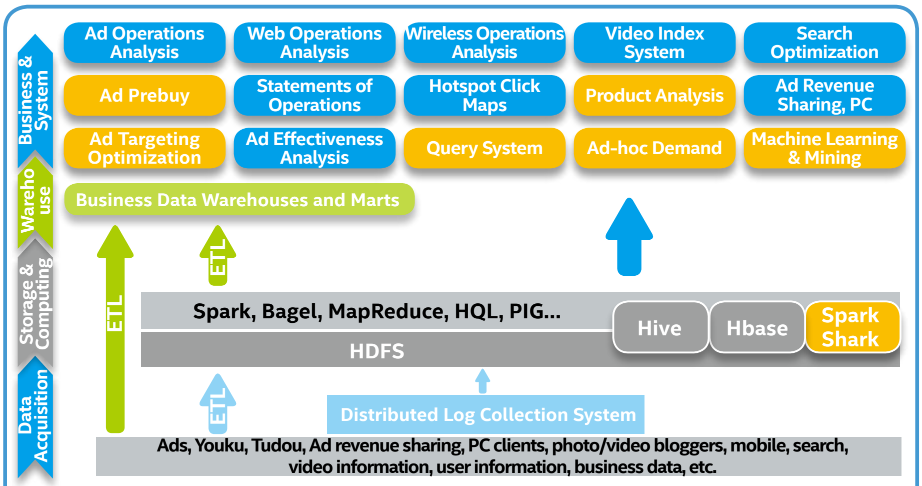
To address this big data explosion, Youku developed a Hadoop cluster. In operation since 2009, it lets Youku efficiently deal with the huge amount of data. During the process, Youku has found that due to the limitations of design patterns, its MapReduce distributed computing framework had an inherent deficiency in processing iterative computation. Explained Fu Jie from Youku's big data team, "In the last few years, we have processed most data with MapReduce and Hive*. We have found that MapReduce was not suitable for all sorts of data processing."

Improving efficiency with enhanced data analysis

As a video-sharing website, Youku has strived to provide unique and personalized solutions for its users. Before realizing this goal, Youku needed to ensure accurate analyses of the huge amounts of data it has generated from its users and videos.

Data is analyzed with the iterative algorithms of machine learning and graph computing. Graph computing is a clustering algorithm applied to analyze the relationships among the videos watched by users. Currently, Youku maintains 200 million videos and around one billion video relationships. (A video relationship is the relevance of one video to another.) Youku computes video relationships with a directed graph computation model. But it was a challenge analyzing such a huge amount of data with a MapReduce solution. In relationship computation, Youku's original solution needed over 80 minutes. In contrast, with the same sample dataset, Spark only needs five or six minutes, which is a cluster and scale-out solution.

Youku Tudou improves its big data analytics and processes with the Spark/Shark big data computing platform, enhancing video-sharing experience by enabling targeted video recommendation application for users



Business & System
Ad Operations Analysis
Web Operations Analysis
Wireless Operations Analysis
Video Index System
Search Optimization
Ad Prebuy
Statements of Operations
Hotspot Click Maps
Product Analysis
Ad Revenue Sharing, PC
Ad Targeting Optimization
Ad Effectiveness Analysis
Query System
Ad-hoc Demand
Machine Learning & Mining
Warehouse
Business Data Warehouses and Marts
Storage & Computing
ETL
ETL
Spark, Bagel, MapReduce, HQL, PIG...
Hive
Hbase
Spark Shark
HDFS
Data Acquisition
ETL
Distributed Log Collection System
Ads, Youku, Tudou, Ad revenue sharing, PC clients, photo/video bloggers, mobile, search, video information, user information, business data, etc.

Architecture Diagram for Shark/Spark Deployment in Youku and Tudou

Spark/Shark

- Spark is an open-source data analytics cluster computing framework that can make data analyses faster. Spark provides a new cluster computing framework, and is designed for certain tasks in cluster computing (i.e., the tasks which require data to be reused, such as iterative computing) in parallel operation. To optimize these tasks, Spark introduces in-memory cluster computing that allows data to be loaded into a cluster's memory, shortening access delay.
- Shark is a data warehouse on Spark. It is compatible with Hive, a data warehose system for Hadoop, and provides high performance Hive queries.

In another application involving TOP-N computing, Youku found that Spark needed only 7.5 minutes and MapReduce needed nearly 30 minutes to analyze the same data set.

This benefit comes as a result of iterative computing under the Spark framework. When faced with complicated tasks such as interactive query and video stream online processing, the MapReduce framework equipped by the Hadoop cluster was less efficient. In contrast, the in-memory computing framework of Spark/Shark was well-suited for various iterative computations and interactive data analyses, since it can directly input the results into the memory (called Resilient Distributed Datasets, RDD), which can be read upon the next usage, saving much disc I/O time and greatly improving efficiency.

Enhancing the Spark/Shark framework setup

Since the graph computation is a CPU-intensive workload, Youku tested the Intel Xeon processor E5-2600 product family and found that it was well suited for their application. Youku has now made the processor family part of the standard platform for its Spark/Shark-based implementation. Intel engineers have been working with teams in Youku to develop their solution and to help them optimize their application performance on the Intel® architecture platform.

Through a video-sharing service called Topic*, Youku hopes to discover its users' interests by analyzing related information among the huge number of user-uploaded videos (more than 1.7 million). Using an in-house-developed, N-weight-based computation model, Youku was only able to achieve two-degree relevancy. The higher the degree of relevancy, the more accurate the prediction of user interests. After implementing a Spark/Shark solution, Youku improved the relevancy up to three to four degrees.

Initially, the Spark/Shark-based solution required 40 hours to compete a computation. Youku improved performance significantly by implementing Intel® Math Kernel Library (Intel® MKL) into its solution. Intel MKL takes advantage of the increased core counts, wide vector units, and varied architectures of modern processors to use a carefully optimized computing math library designed to harness full processor potential. Intel MKL includes highly vectorized and threaded linear algebra, fast Fourier transforms (FFT), vector math, and statistics functions. Through a single C or Fortran* API call, these functions automatically scale across processor architectures by selecting the best code path for each. After implementation of Intel MKL, Youku reduced the computation time to less than three hours. The higher degree of relevance and significantly enhanced performance help Youku bring better user experience to customers using Youku services.

Innovating better services

As one of the first companies in its sector to commercially apply the Spark/Shark framework, Youku has adopted an ideal, groundbreaking solution. It hopes to continue working with Intel during the optimization of the overall plan and the upgrading of its hardware and video-sharing system.

Find the solution that's right for your organization. Contact your Intel representative, visit Intel's Business Success Stories for IT Managers (**www.intel.com/itcasestudies**) or explore the Intel.com IT Center (**www.intel.com/itcenter**).

This document and the information given are for the convenience of Intel's customer base and are provided "AS IS" WITH NO WARRANTIES WHATSOEVER, EXPRESS OR IMPLIED, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT OF INTELLECTUAL PROPERTY RIGHTS. Receipt or possession of this document does not grant any license to any of the intellectual property described, displayed, or contained herein. Intel® products are not intended for use in medical, lifesaving, life-sustaining, critical control, or safety systems, or in nuclear facility applications.

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.

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

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.