

# Taobao.com Runs China's Busiest Data Warehouse on 20 Node Database Cluster



Taobao.com  
Hangzhou, China  
www.taobao.com

## Industry:

Media & Entertainment

## Annual Turnover:

US\$14.7 billion

## Employees:

More than 2,000

## Oracle Products & Services:

Oracle Database

Oracle Real Application Clusters

*“Oracle Real Application Clusters supports large-scale data processing and can scale to meet our rapidly growing requirements. It provides a powerful and reliable platform on which to run our Oracle data warehouse.” – Wang Hai, Senior Supervisor, Taobao.com*

Founded in 2003 by Alibaba.com, Taobao.com is Asia's largest online shopping Web site. The company engages in consumer-to-consumer (C2C) and business-to-consumer (B2C) trading and has a 70% share of the online market in China. More than 110 million registered members purchased goods and services worth US\$14.7 billion (RMB 99.96 billion) in 2008. More than 10 million cell phones and over 140 million items of clothing were sold through the site in that year alone.

Taobao.com has used Oracle Database and clusterware products since 2004, building an enterprise data warehouse to process and analyze business and customer data.

The company recently upgraded from Oracle Database 10g with Real Application Clusters to Oracle Database 11g with Real Application Clusters. As part of the upgrade, Taobao.com deployed a 20-node database cluster to improve data processing speed. This has enabled the company to cut the time required to run reports from days to minutes and, in some cases, to near real time. The 30-terabyte data warehouse processes hundreds of millions of transactions daily, making it one of China's busiest and most powerful applications.

“The Oracle data warehouse must process new and existing data every hour,” said Chen Jiping, chief database administrator and senior technology expert, Taobao.com. “Oracle Real Application Clusters satisfies our requirements for reliable, efficient mass data processing, which ensures that we always have the information we need to improve the management of our business.”

## Customer Data Is Key to Improving Competitiveness

The e-business market in China is maturing rapidly, as evidenced by Taobao.com's success. Many e-businesses are now competing

**Key Benefits:**

- Processed 30TB of data with ease on a daily basis
- Reduced data processing time from hours to minutes
- Enabled 500 ETL tasks to be completed in 8.5 hours
- Provided staff with analysis that gave them a deeper understanding of customers, enabling them to offer personalized service that enhanced brand loyalty
- Accommodated rapid business growth with scalable solution, which enables nodes to be added when extra processing power is needed

on price, prompting Taobao.com to switch its focus to improving customers' shopping experience and strengthening their loyalty.

To do this, the company must offer more personalized service, a task that can only be achieved if it has a deep understanding of its customers and their shopping profiles. This vital information was available in various forms in the data warehouse; the challenge was how to unlock and deliver the information in ways that would enable the entire organization, from senior managers to customer service staff, to improve the way it engaged with customers.

**Meeting the Challenge of Mass Data Processing**

Taobao.com's data warehouse contains 30TB of information, and almost all of it needs analysis on a daily basis. For example, users are invited to comment and rate the stores that sell goods on the Taobao.com site. These comments and ratings need to be monitored daily for misleading and defamatory statements, which ensures the integrity of the feedback and rating system. Taobao.com uses the data warehouse to run reports that consolidate the day's comments and ratings for staff to review.

Another feature of the Web site is i-Taobao, a service that recommends products, services, and stores to customers based on their historical transactions. For example, if a user has purchased items for babies in the past, the site may refer them to stores that specialize in baby clothes or toys, or showcase new products similar to those bought by the customer in the past. However, to ensure the right recommendations are made at the right time to the right members, the data warehouse must be able to process and analyze data at lightning speed. With more than 110 million members and over 120 million products on offer, this is no easy task.

**Scalability Enables High-Speed Data Processing**

In 2009, Taobao.com increased its 12-node data warehouse environment to 20 nodes. Data scattered in different systems was consolidated in the data warehouse. This data includes access rates, transaction numbers, and call center queries. The information is cleaned, filtered, and consolidated into data marts, where it is displayed as indexes and statements. In this way, staff can view accurate, up-to-date analysis on customers' browsing behavior, purchasing patterns, and product preferences.

With this information at its fingertips, Taobao.com has developed a deep understanding of its customer base, enabling the

organization to design strategies to offer personalized services and improve the range and quality of products on offer.

“There is a significant difference in the efficiency of parallel and non-parallel processing,” said Chen. “Oracle Real Application Clusters offers powerful parallel processing capabilities that ensure we can process large amounts of data concurrently.”

This parallel processing capability is particularly useful for the i-Taobao service. After upgrading to the 20-node Oracle database cluster, the time needed to process data for customer recommendations has been reduced; it now takes between 30 minutes and two hours to crunch data.

“The data warehouse also produces up to 400 reports,” said Chen. “These reports look deceptively simple to the users, but they are the result of complex data processing. The information they contain is critical to the success of our business—there is no way we could do this manually or by using less sophisticated databases.”

### **Easy Scalability Supports Long-Term Growth**

The use of Oracle technology ensures the data warehouse can be easily expanded to accommodate Taobao.com’s growth and the corresponding explosion in data volumes. The company implemented a four-node database cluster in 2004 then increased the cluster to 12 nodes in 2008 and to 20 nodes a year later. Taobao.com has experienced a reduction in the time needed to process information, despite the massive increase in data volumes. In some cases, what took several hours in the past can now be completed in minutes.

“The linear scalability of Oracle Real Application Clusters is of great importance to us,” said Chen. “It allows us to boost computing capacity simply by increasing the number of nodes. On some jobs, the processing period for the same amount of data has been halved. This flexibility ensures our business can keep pace with changing customer and market demands.”

For example, following a period of rapid growth, Taobao.com found it could not finish running the data it needed to support core businesses using the four-node database cluster, despite an 8.5-hour processing schedule. After expanding to 12, then 20 nodes, the company can now complete up to 500 extract, transform, and load (ETL) tasks in the same amount of time.

### Why Oracle?

Taobao.com is a dynamic online business, with a large volume of daily transactions and an increasing customer base. As such, the company was looking for a robust, reliable, and scalable clustered database product to underlie its data warehouse. Oracle Real Application Clusters was selected for three advantages.

First, the software has parallel processing capabilities, which supports dynamic querying and analysis and meets Taobao.com's demands for large-volume processing in a timely manner.

Second, the linear scalability of Oracle Real Application Clusters appealed to Taobao.com, as it would allow the company to use low-cost commodity servers and add nodes only when extra processing power was needed. This would help constrain costs, which is particularly important in a tough economic environment. The flexibility to add servers based on demand also eliminates the need to split the data warehouse into smaller data marts to ensure stable system performance.

Third, Oracle offered tools that would simplify maintenance and reduce management costs. Oracle Automatic Storage Management (ASM) enables administrators to automatically create and delete data folders and improve storage efficiency, making it easier to maintain the data warehouse.

### Implementation Process

Taobao.com built its first Oracle data warehouse in 2004. It was initially run on a single server, but this proved untenable when the company experienced rapid growth. The data warehouse environment was expanded to four nodes, then 12 nodes, and finally 20 nodes and runs on Linux servers.

*Founded in 2003 by Alibaba.com, Taobao.com is Asia's largest online shopping Web site. The company engages in consumer-to-consumer (C2C) and business-to-consumer (B2C) trading and has a 70% share of the online market in China.*