High Performance Scoring with Oracle 10.2 Data Mining

Scoring-in-the-large

An Oracle White Paper
April 2006

High Performance Scoring with Oracle 10.2 Data Mining

High Performance Scoring with Oracle 10.2 Data Mining – Scoring-in-the-large

EXECUTIVE OVERVIEW

Once the data mining models are created, the fun begins. Model deployment, or enabling the use of models to produce results, should be easy…and with Oracle Data Mining in-database model building and scoring…it is!

Organizations and businesses are accumulating mass quantities of data to meet regulatory compliance requirements and support corporate-wide business intelligence. Corporate management needs to derive more value from their data and avoid the terms "data graveyard" or "tax" applied to their data warehousing projects. The corporate data warehouse and operational data stores and rich resources for gaining insight into one's business and leveraging the knowledge contained in the data to make predictions, find unexpected relationships among components, or better understand one's customers.

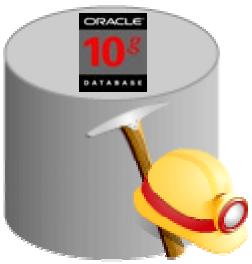
Corporate management is coming to understand the value that data mining can bring to an organization, yet companies often fall short of realizing data mining's potential return on investment because of difficulties in deploying models to applications or reporting systems. In this white paper, we discuss how Oracle Data Mining simplifies model deployment and increases the ease with which data mining results can be incorporated into applications and reporting systems through the use of SQL.

Oracle Data Mining, however, takes several steps forward by providing a high-performance scoring engine that provides scalability at unprecedented levels using modest off-the-shelf hardware. In-database mining provides security, backup and recovery, and minimizes the IT effort to support data mining-based initiatives.

Through an internal benchmark, we show a large-scale customer scenario involving the scoring of 100 models over 100 million customers, which can be done in an overnight window, ready for use the following day. This is all possible with an inexpensive 4-CPU Linux box.

ORACLE

BUSINESS INTELLIGENCE



INTRODUCTION

Organizations and businesses are accumulating mass quantities of data to meet regulatory compliance requirements or support corporate-wide business intelligence. Unfortunately, the corporate data warehouse can also be perceived as a data graveyard, or a tax on the corporation, instead of a rich resource for gaining insights into one's business and leveraging the knowledge contained in the data to make predictions, find unexpected relationships among components, or better understand one's customers.

There is a critical need to derive value from corporate data and data mining is a key technology to make this a reality. Whether the industry is retail, healthcare, financial services, or communications, the need for customer analysis, root cause analysis, product cross-sell and upsell is pervasive. As customer bases increase into the tens of millions and purchased demographic data is available for 100s of millions of potential customers, the ability to use this data effectively through efficient data mining requires revolutionary approaches, a paradigm shift.

The paradigm shift is from considering data mining as an ancillary activity, conducted in separate software on remote machines, to bringing data mining to where the data is already, in the database. By mining the data in the database, within the same process space and using similar techniques as high-performance queries, model building and scoring achieve orders of magnitude performance improvements.

In a fast-paced business climate, it is critical to produce and consume results with minimal delay; otherwise, the results lose their value. Historically, model deployment was costly and risky to introduce into a working process each time a new model was produced. Today, the cost and risk is greatly minimized, if not eliminated, through in-database scoring. Part of any automated process is the need to refresh models, i.e., to rebuild the models on more recent data, and then to redeploy them. In the database, this is immediate since the scoring query remains the same; only the underlying model is replaced in the database.

So, what is model deployment? Model deployment is the process required to use a model in a target environment. This may be as simple as (i) extracting the model details, e.g., the rules of a decision tree or attribute importance ranking, to produce reports, (ii) scoring data using the model in the database where it was built either

for batch or real-time results, or (iii) moving the model from the system where it was built to the system where it will be used for scoring. Model deployment can also involve altering applications or operational system to incorporate data mining results as part of their normal behavior; however, this should be a one-time investment, not a recurring theme.

SCORING DATA USING ORACLE DATA MINING

The "old school" of thought viewed data mining as a separate activity: get an extract of the data, prepare it, build the models, send the models to developers to code into applications or score data writing the results to flat files, and then try to get the results into the operational systems. This process could take weeks or months. From the description, it is clear there were many moving part with plenty of opportunity for hand-off errors, miscommunication, and security lapses.

The "new school" of thought reflecting this paradigm shift prepares the data in the database, using high performance query processing and analytics available in the database. It produces the models in the database, which remain as database objects that can be used in place or exported to other database systems. These models can be immediately used as part of operational systems through the execution of SQL statements to produce individual scores or entire batch scores. For example, the following SQL produces a prediction and probability for a specific customer from the CUSTOMER table. This type of a result could be used in a call center application to make offer recommendations that the call center representative can provide to the customer.

```
select prediction(DM_MODEL using *) offer,
       prediction_probability(DM_MODEL using *) probability,
from CUSTOMER
where cust_id = '100015'
```

Alternatively, the entire CUSTOMER table could be scored, even against multiple models using a straightforward SQL statement. These results can be consumed as part of a stream (without incurring the cost of materialization), or written to a table for subsequent use.

```
select cust_id,
       prediction(DM_MODEL_1 using *) offer1,
       prediction_probability(DM_MODEL_1 using *) probability1,
       prediction(DM_MODEL_2 using *) offer2,
       prediction_probability(DM_MODEL_2 using *) probability2,
from CUSTOMER
```

A DEMANDING SCORING SCENARIO

Consider a situation where a company purchased from a third party 100 million person records containing demographic information. Their goal is to determine which of those persons to target for any of 100 offers. As new data arrives from previous offer campaigns, models are refreshed, i.e., rebuilt and redeployed, and the potential customers are scored again. New offers are frequently being added and removed from the mix.

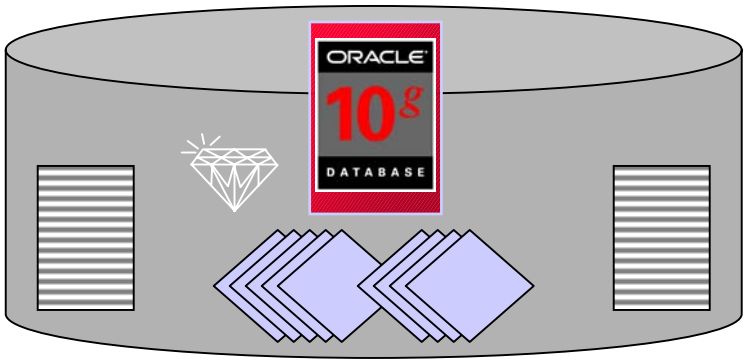
The goal is to score each potential customer using each of the models and store the results in a database table for subsequent use by applications. The company needs to produce the scoring results in an overnight time window.

LEVERAGING AN OFF-THE-SHELF SOLUTION

Most IT organizations have limited hardware and human resources to devote to any new initiative, including data mining. As such, a solution that enables efficient management of models and high performance scoring on inexpensive hardware can enable data mining to be a must-have component to the business process – not a luxury. Add in that in-database mining allows the data and models to *always* be managed as database objects, gaining the benefit of security, backup and recovery, and minimizing the overall IT effort to leverage data mining. Fewer moving parts means less complexity and maintenance.

SCORING-IN-THE-LARGE BENCHMARK DESCRIPTION

To highlight the performance capabilities of Oracle Data Mining, we have selected 100 million cases with 250 attributes predicting one binary target. This data exists in Oracle 10g Database. We built 100 classification decision tree models in the database to be used for in-database scoring. The resulting scores are materialized in a database table for application use or analysis.



100M cases
250 attributes
1 binary target

100 Decision Tree Models
In-database scoring

Scores from 100 models
on 100M cases

Figure 1: Scoring-in-the-large benchmark definition

TEST ENVIRONMENT CONFIGURATION

The performance tests were run on a fairly common machine available at or easily within budget of most companies: a machine with 4 Intel Xeon processors at 3.4 GHz, 4 GB RAM, 8 GB swap space, and running the Linux operating system. The disk storage involved 115 GB of a 4.8 TB EMC Clariion CX700 box supporting a storage area network (SAN).

TEST RESULTS

With the power of in-database scoring no unusual system configuration is required. IO usage, CPU usage, and parallel computing are all automatically optimized by the database. For this benchmark, all processors were kept fully utilized during scoring. A single pass through the data occurred, scoring all 100 models for each case (record).

# of Rows	Time in Minutes	Scores per second
25M	80	520,500
50M	159	524,328
75M	238*	525,210*
100M	316	527,649

* Interpolated

To score the 100 million customer records with 100 models took 5 hours and 24 minutes. From the table above, note that the number of scores per second actually increased as the number of rows were processed, indicating a sublinear scalability – scoring actually got faster! This is depicted in Figure 2 – ODM's in-database scoring scales sublinearly with the number of cases on a basic 4 CPU Linux machine. If the business requirements had a shorter time window or the processing of more customers, additional low-cost machines could be employed. For example, with 8 CPUs, we could cut scoring time in half, completing the 100 million by 100 models scoring task in about two and a half hours. This scalability can be accomplished easily using Oracle's Real Application Clusters (RAC) / GRID computing capability.

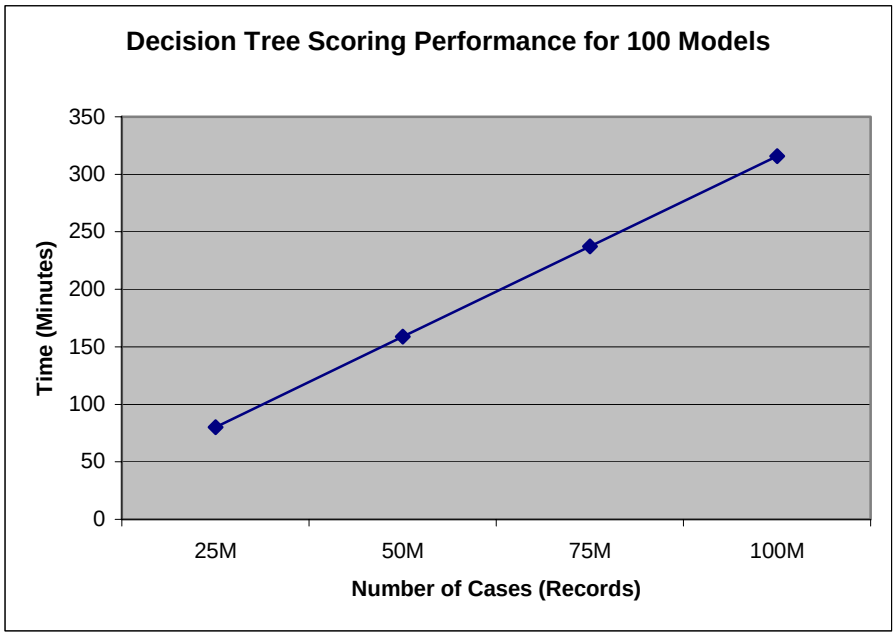


Figure 2: Scoring-in-the-large benchmark results

This capability, via in-database mining, changes the landscape of what is possible for applications and business intelligence in general to use data mining.

Oracle Database provides mechanisms to allow partitioning and scheduling of workloads, e.g., RAC, database scheduler, etc. As a database technology, data mining activities can be treated and managed as any other database workload.

CONCLUSION

The need for scoring-in-the-large is already here. Companies have the data but may not feel they can adequately leverage it to gain insight and knowledge through the use of data mining. In this white paper, we have shown how data mining models can be deployed efficiently and easily as part of the database environment; moreover, that scoring large volumes of data can be done in very tight time windows using relatively inexpensive hardware.

IT organizations and DBAs can apply their Oracle Database skills for database management to data mining objects, including access control, backup and recovery procedures, and installation and maintenance.

MORE INFORMATION

Oracle Data Mining (ODM), an option to Oracle Database 10*g* Enterprise Edition, enables companies to extract information efficiently from large databases and build integrated business intelligence applications. Data analysts can find patterns and insights hidden in their data. Application developers can quickly automate the extraction and distribution of new business intelligence—predictions, patterns and discoveries—throughout the organization.

http://www.oracle.com/technology/products/bi/odm/index.html

Oracle Data Mining - White Paper (PDF, 1.5MB)

Oracle Data Mining Overview (updated for Oracle Database 10*g* Release 2) - Presentation (PDF, 1.5MB)

Product Discussion Forum

http://forums.oracle.com/forums/forum.jspa?forumID=55

ORACLE

High Performance Scoring with Oracle 10.2 Data Mining
Version: 1.1
April 2006
Author: Mark Hornick

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
www.oracle.com

ORACLE
BUSINESS INTELLIGENCE