Oracle AI & Data Science Blog

Oracle AI

Machine Learning Autonomously 7.4 7)

July 26, 2019 | 12 minute read

Victor Lu

From Tuning Manually to Optimizing Data 7.4 7)

When will the query finish?

At 9:00pm, I was still in the office. I wanted to run a data quality report and print it out to show my boss the next day.

My wife called. It was the 3rd time she called; she was upset because I promised to go home by 7:00pm. I canceled the query and submitted it one last time.

But this time, I had given up any hope of having data ready for statisticians one day ahead of schedule. Instead, I created a ticket for the database administrator to find out why my new query did not finish in time.

That was in 1995. I was a database analyst without full access to the database backend, and it was a mystery to me why my database report did not finish in time. I wished to have the knowledge and power to analyze and control the situation myself.

I was lucky -- I became a database administrator not long after that. I began to manage multiple database platforms: Oracle, Sybase, and Microsoft SQL Server. It was fun to be the king of the enterprise database world, dictating what can or cannot be run.

Gaining More Insights into Performance

Oracle databases are usually the largest, most scalable databases, but they're also the easiest to tune -- I can gather information from v$session and v$session_wait database views. These views expose an infrastructure called Oracle Wait Interface that takes the guesswork out of performance tuning.

Even if it is relatively easier to identify performance problems in an Oracle database compared with Sybase and SQL Server, database performance is still unpredictable. Some in the industry have claimed that tuning is a nightmare and that auto-tuning is wishful thinking.

Finally, Oracle decided to take on the challenge and get rid of its rule-based optimizer – the transition from a DBA manually tuning queries to a database automatically tuning SQL queries.

In the early days of database upgrades to a version that only has a cost-based optimizer, there were Oracle DBA's and developers who got frustrated trying to figure out the best way to collect database statistics and use Oracle tools to stabilize performance. (Even nowadays, there are still database administrators that do not collect database statistics and use all the tools properly.)

Oracle documentation has a very good analogy for the cost-based optimizer: an online trip advisor.

Let's say a cyclist wants to know the most efficient bicycle route from point A to point B. The advisor picks the most efficient (lowest cost) overall route based on user-specified goals and the available statistics about roads and traffic conditions. The more accurate the statistics, the better the advice. For example, if the advisor is not frequently notified of traffic jams, road closures, and poor road conditions, then the recommended route may turn out to be inefficient (high cost).

Collecting stats to cover all scenarios is the main driver behind the success of a cost-based optimizer. Oracle has embarked on this since Oracle 7 in 1992.

Oracle Optimizer did not stop at the database tier. With the introduction of Exadata, Oracle optimization goes deep into the storage/InfiniBand network layer. Smart Scan allows most of the SQL processing to happen in the storage tier instead of the database tier, which dramatically improves query performance. It reduces the volume of data sent to the database tier, thereby reducing CPU usage on database nodes. Similarly, Oracle Big Data SQL also has a smart scan optimization feature, which enables organizations to immediately analyze data across Apache Hadoop, Apache Kafka, NoSQL and Oracle databases.

The Journey to Self-Driving Technology

Compared with LIDAR for self-driving cars, Oracle's machine learning-driven database is not popular news. Yet, this technology made it possible for a database to perform well through learning from its own experience, running a SQL query, and deciding what to change.

Elon Musk recently commented, "LIDAR is a fool's errand. Anyone who relies on LIDAR is doomed. Expensive sensors are unnecessary."

Tesla is planning to use only cameras and radar without LIDAR for Tesla's self-driving cars. Whether Elon will succeed is still to be watched, but his remarks do show that Elon and Larry had a similar vision – that artificial intelligence and machine learning do not need to rely on fancy hardware or even a sophisticated machine learning framework to get implemented.

Life was not easy for Oracle DBA's, and neither was it for the engineers testing self-driving cars. I have full respect for all the people who worked through the early days, when you couldn't use database rule-based hints to force SQL query to use an index lookup over a full table scan.

After over 2 decades of improvements, the Autonomous Database is finally here -- unlike self-driving cars, which are still a dream for the future.

Fast-forward to OpenWorld 2017 – right after Oracle unveiled the world's first Autonomous Database Cloud. Suddenly, many database vendors claimed that their managed database platforms were autonomous, truly self-driving as well. Whenever I talked about these other "autonomous databases" with Oracle database administrators, we all laughed...

Scale Machine Learning with the Autonomous Database

Now back to the present day, 2019: I was talking to a developer about his data science project where he brought all the data into Python pandas dataframe. The real-time analytics and beautiful graphics caught my attention.

Because of my "occupational hazard" as a longtime DBA, I was curious about how he processed everything in an AWS EC2 instance that was not even as powerful as a low-end laptop. He answered, "Because the source data is less than 100 Megabyte."

That made me wonder, "What if the source data is more than 100 Terabyte?"

While there are use cases where you only need a small dataset, the growing volume of data has caused data scientists to use sampling techniques that could reduce the accuracy of machine learning models.

Instead of risking significant sampling bias or information loss, why not bring the machine learning algorithm to the data?

Machine Learning, Right Where You Need It

Actually, Oracle has done that for years. It is called Database Advanced Analytics (OAA), and the components are Oracle DB + Oracle Data Mining + Oracle R Enterprise.

The term "deep learning" is used in conjunction with neural networks. Starting with Oracle Database 18c, the Neural Network machine-learning model was introduced. You can use the model to label new data, which is something that is typically required in your front-end applications, web-based applications, or your back-end or batch applications and processes. Any programming language or framework that can call SQL can now run the in-database Neural Network model by using a simple SQL command.

One interesting experiment in the database/big data world right now is the use of innovative hardware accelerators such as Graphics Processor Units (GPUs). They run SQL queries on multi-billion-row data sets, and they enable machine learning algorithms and code to run on data directly within the database. These GPU databases, even if they become mature enough and available on Oracle Cloud Infrastructure, will only complement Oracle in-database analytics. Just like their slower yet similar predecessors (such as Amazon Redshift and GreenPlum), they are not scalable as an OLTP database engine.

As a multi-model database, Oracle in-database machine learning can support the majority of use cases in every industry. With Hybrid Transaction/Analytical Processing (HTAP) capabilities in the same database, Oracle database can analyze data "on the spot."

Oracle database, to this date, is still the ONLY DATABASE that is best at handling high-volume OLTP transactions and offers an in-database analytics platform. Only the self-driving Oracle Autonomous Database can bring machine learning as close to the core business workflow as possible.

Now that Oracle made it a breeze to create a production-ready autonomous database and implement SQL queries with minimal user interaction, there is really no reason not to get started with Database Advanced Analytics.

To learn more, check out these Oracle documentation chapters on Machine Learning models that are already available to use as SQL functions on Oracle database.

How Oracle Database Sets the Stage for Data Science Projects

Even if your company is not ready to try in-database machine learning, Oracle database can jumpstart successful data science and machine learning projects.

Oracle is known to be far superior as an OLTP database. It was also positioned highest for its ability to execute and completeness of vision in Gartner's 2019 "Magic Quadrant for Data Management Solutions for Analytics" report. Compared with the 2018 report, Oracle is expanding its lead on Gartner's vertical axis over Microsoft and Amazon Web Services.

This unique and mature platform is not easily imitated. Without rewriting the database engine from scratch, it will be very hard for other platforms to handle high-volume OLTP transactions because of architecture defects such as "Auto-vacuum" or "reader block writer, writer block reader."

After several years of the "NoSQL revolution," more and more enterprise users and startups have realized that SQL databases have a major advantage over NoSQL databases in many areas. A SQL database that is self-driving, self-securing, and self-repairing, as well as highly scalable vertically and horizontally, is critical in any data science project.

Here are some reasons why the "Ready to Work" Oracle databases can lead to successful data science projects:

- **You already have data.** Data is the fuel of the Fourth Industrial Revolution. Google and Facebook manage the world's largest collection of data. Even if they are hit with $8.8 billion in lawsuits on day one of the new privacy regulation GDPR, they will still have an advantage when it comes to using personal data to feed their data science platform. What about your organizations – banks, telecom, manufacturers, utilities, airlines, retailers? With Oracle owning almost half of the world's database market share, I bet that Oracle databases contain the "New Oil" that many organizations own.

- **Historical data gives you the competitive advantage.** These corporate knowledge bases cannot be created or replaced – if you have every played in the stock market, you will understand how historical data teaches you the up and downs of the stock market and related historical events. Many customers have similar historical data. In supply chain analysis, it is possible to correlate your historical item data, past requirements, and actions that have been taken, so that you can easily explain why an inventory issue exists.

- **Strategic data acquisition, NOT from scratch.** Most of your Oracle databases already contain data about your products and your customer. If you launch a new product or enter a new market, it is much easier to analyze the data you already have, learn from past success and mistakes, and make strategic plans on how to correlate data from different data sources. With this virtuous cycle of machine learning, you can gain momentum quickly in any pilot data science projects and leverage in-house Oracle platform development expertise as an integral part of your project. Successful pilot projects will result in more data, and better data, for the next

development expertise as an integral part of your project. Successful pilot projects will result in more data, and better data, for the next iteration of your data science/machine learning project efforts.

- **Your data in an Oracle database are clean.** The first step in your data science or machine learning workflow is cleaning data. Your Oracle database, whether it's an OLTP or data warehouse, allows you to define and enforce the most comprehensive type of data integrity rules, already normalized for proper storage and data retrieval. Data in existing Oracle data warehouses have already been used extensively in your corporate analytic processes. These data are well understood, pre-processed, and clean. Therefore, you are much more likely to go directly into the data analysis phase if you leverage these data as part of your new machine learning efforts.

- **Regulatory compliance and data governance**. After years of supporting organizations to meet ever-increasing regulation requirements, Oracle database systems have already been battle-tested with many industry best practices. They can support regulations such as SOX, HIPPA, and the new GDPR data privacy rules. As a matter of fact, Oracle was named a leader in KuppingerCole's latest Leadership Compass: Database Security. Compared with building new data lakes and making uncontrolled, massive data repositories, starting your enterprise data science project with your well-governed Oracle database reduces regulatory compliance risks. These efforts result in better data governance, better master data management, reduced data duplications, and better support in the never-ending quest for a single customer view.

- **You already have the right development team**. Similar to scientific research, the subject matter experts in your company know your business and your data the best. These people already use the applications to run your business; they understand whether the data are stored properly, clean, and ready for further analysis. Your current Business Analyst/Application Development team has the skills needed to support business requirements. Oracle database developers are either already skilled or can be easily trained to develop solutions in uses cases for multi-model deployment in the same Oracle database. With proper training and an open mind to all the latest machine learning technologies and development methodologies (Think Agile, CI/CD, and DevOps), they can work with your data science team to launch a data science project that will give a quick return on investment (ROI).

- **Enterprise Architecture is not a new word to your enterprise architects.** For many of the Fortune 500 companies, enterprise architects have worked with data in your Oracle databases for decades, trying to put information in every person's hands so they can use data better. These enterprise architects understand your business, the data you already have, and the strategic data you can acquire from your new IoT, Robotic, or mobile/5G initiatives. They combine business strategies with technology strategies to help transform your business and spur added innovation using IT systems and good architecture principles. Why should enterprise architects turn to Oracle for database solutions? Oracle not only offers the best on-premise database, but it's also the most innovative Database as a service (DBaaS) leader.

With Oracle databases, you can simplify data management, correlate different data sources for deeper insights, and enable your existing teams to innovate through new projects.

Andrew Ng maintains that you're not really ready to be an AI company until you make strategic data acquisitions and gather all your data into a centralized data warehouse. Bring it on then – your AI company already has Oracle databases. Even better, they are becoming autonomous.

Not Everything is Self-Driving, and That is Perfect

Even if Oracle's Autonomous Database is not the best fit for your company's data, there are plenty of Oracle offerings that can better fit your unique needs.

Exploring Microsecond Response Time with TimesTen

Most Oracle-based applications can live with a millisecond response time. However, real-time applications require microsecond query response time. Therefore, Oracle acquired TimesTen, an in-memory database company, in 2005. While working with customers, I achieved query performance close to single-digit microseconds. This was not thinkable, even with the latest release of Oracle database.

There was quite some excitement in the Oracle customer base that year. I got called into numerous customer sessions regarding TimesTen use cases. In 2006, there were still not many well-known solutions as scalable as Oracle database in ANY use case, even for the startup world.

However, one of the customer's requests stumped me. He wanted to embed Oracle database in an application server to store session state information. Data would be partitioned with a replication of session state between the application servers.

Oracle database's footprint was too big for this use case, and it would have been overkill to be a local store for session state. TimesTen was a possible solution; however, it was a new use case that needed to be investigated.

While I was still doing research, Oracle announced their acquisition of Sleepycat, which owns an alternative NoSQL database - BerkeleyDB. The software is embedded in open-source products such as the Linux and BSD Unix operating systems, the Apache Web server, the OpenLDAP directory, and the OpenOffice application suite. Perfect!

Find the Right Fit for Your Data

As matter of fact, there are more and more products in Oracle's ecosystem that are great for a specific use case when Oracle's self-tuning, highly available database might be overkill.

Over the years, I have tried to use Oracle database as a scalable solution for both structured and unstructured data. Although Oracle is a multi-model database, my experience tells me that it is not the best fit for every scenario -- but it should be the centerpiece of any enterprise data science effort.

There are many other Oracle solutions available. MySQL is designed for the web; it scales well and complement Oracle RDBMS very well. Oracle NoSQL database was an early darling of many Silicon Valley startups.

Oracle is one of the largest producers of open-source software in the world, developing and providing contributions for projects including Apache NetBeans, Berkeley DB, Eclipse Jakarta, GraalVM, Kubernetes, Linux, MySQL, OpenJDK, PHP, VirtualBox, and Xen. Oracle Cloud

Infrastructure's core services are built on open-source technologies to support workloads for cloud native applications, data streams, eventing, and data transformation and processing.

The upcoming Data Science Cloud will feature open-source options. Built-in, cloud-hosted Jupyter notebooks enable teams to build and train models using Python. Popular open-source visualization tools like Plotly, Matplotlib, and Bokeh allow you to visualize and explore data. You can launch environments with popular machine learning frameworks like TensorFlow and scikit-learn, or easily install any other open-source package.

This solution will also use native support for most popular version control providers (Github, Gitlab, and Bitbucket) to ensure that all work is synced across the platform. Tight integration with OCI and Oracle Big Data Platform provides data scientists with self-service access to scalable compute, so that they can get to work quickly.

On Oracle Data Science Cloud, you can deploy a function or model as an API to encapsulate data science work and expose it to other team members and applications. With APIs, you can:

- Provide predictions as a microservice to a larger application
- Provide inputs to visualization software like Tableau or Bokeh
- Share your model with other analysts and data scientists.

Looking for scalability and security? Spin up containers on Oracle Cloud Infrastructure to tackle analyses of any size leveraging Kubernetes. You can use secure credentials to safely access data in Oracle Object Storage Cloud.

Oracle Data Catalog enables users to discover, find, organize, enrich, and trace data assets from multiple sources. You can put your data to work to create a unified data warehouse, including "ready to go" data sources from cloud SaaS databases.

For analytic workloads over large and complex datasets, SNAP, an Apache Spark™ native business intelligence platform, brings sub-second query response times that can be deployed over enterprise data warehouses, data lakes, and IoT analytics.

Although Oracle database is the world's only real autonomous database for OLTP and Analytics workloads, there are new cloud services introduced all the time that give the right hammer for your nail. This way, the focus is not on the tools you use, but on how you can drive business innovation and revenue growth –and that is PERFECT!

Victor Lu
Oracle Solution Specialist