IBM Analytics
White paper

Machine learning scoring on premises or in the cloud? You decide.

IBM measures the impact of these two scoring approaches

Organizations across all industries realize that their data holds the key to future business success and that machine learning is vital to unlocking the value of that data. As organizations implement machine learning, they must consider the best approach: on premises, in the cloud or a blend of both. When some of your enterprise data within business-critical applications resides on the IBM® Z® platform, you can leverage IBM Machine Learning for z/OS®, colocating your machine learning processes with your data to improve efficiency.

The impact of scoring execution time

Machine learning is about building models and then deploying them to make predictions. Scoring can then leverage the deployed model to make a prediction based on new data. When incorporating scoring within a production operational application, execution time is critical. With significant transaction volumes, just a few more milliseconds can have millions to tens of millions of dollars of impact on revenue. Colocating the scoring process with your transactional processes on the IBM Z platform can minimize the execution-time impact.

To determine the impact of machine learning scoring on service-level agreements (SLAs), IBM compared scoring locally, using IBM Machine Learning for z/OS, to scoring in the cloud. It examined the throughput and response times of the machine learning scoring services and how these metrics impact SLAs.

IBM selected a banking customer retention use case, since many traditional banks are losing customers to financial technology (FinTech) companies. FinTechs are perceived by many people to be more efficient and better at leveraging new technology, such as cloud, to provide innovative financial solutions. To predict whether customers are likely to take their business to another bank like a FinTech, IBM used sample data on IBM Z to build and train a customer churn model.

Model building can be accomplished by keeping the data locally on the IBM Z platform or moving it to a data lake. Organizations can use either approach; there are IT and business rationales for each. In this example, the models were built on the IBM Z platform, using IBM Machine Learning for z/OS, to reduce the time, expense and security breach risk of moving large quantities of data to a data lake.

Once a model is built and deployed, the scoring service can be called during any real-time interaction with a customer through a mobile app, web app, ATM transaction, teller transaction or in a batch process. The scoring service leverages a simple RESTful API request. This interface uses various demographic and other data about a customer and responds with a prediction about whether that customer is likely to churn. By exploiting the scoring services' predictions within the operational applications, proactive measures can be taken, such as targeting special offers or providing personalized services to help retain at-risk customers.

Throughput, response times and consistency matter

A well-utilized scoring service can be called many thousands of times per minute. It must support high throughput, fast response times and consistently meet SLAs. Since the application resides on the IBM Z platform, logic suggests that scoring on the same platform will provide greater throughput and faster, more consistent response times.

IBM drove the performance test using Apache JMeter to confirm this logic. The performance results were stored in an InfluxDB database and rendered on a Grafana dashboard. The two metrics measured were the number of executions per second and the average response time.

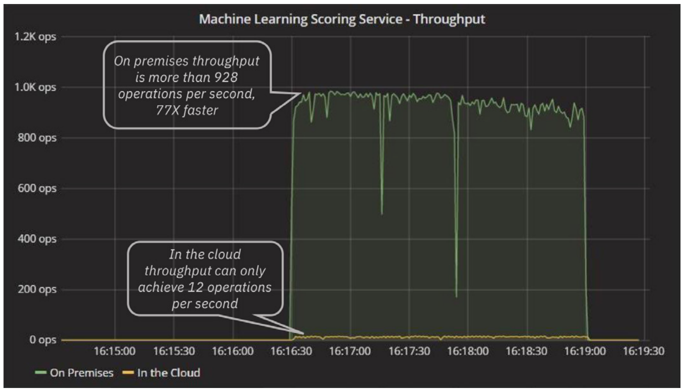

Figure 1: Screen shot showing the throughput experienced using the on-premises approach

The IBM Z on-premises approach, using Machine Learning for z/OS, achieved greater than 77 times more throughput, averaging more than 928 round-trip requests per second. The cloud approach could only achieve 12 round-trip requests per second, as shown in Figure 1.

The difference in response times is just as dramatic. The on-premises approach measured less than 1 millisecond or 83 times faster, on average, than the cloud approach, as shown in Figure 2.

Response-time SLAs are typically based on the percentage of responses executed within specific time limits. It's important to observe that the on-premises response times are also much more consistent. This variability is obvious in Figure 2, and further illustrated in Figure 3. These figures show that the on-premises 99th percentile responses are only 2 milliseconds or less, while the in cloud 99th percentile shoots up to 274 milliseconds. This significant variability presents challenges in both writing and meeting SLAs.

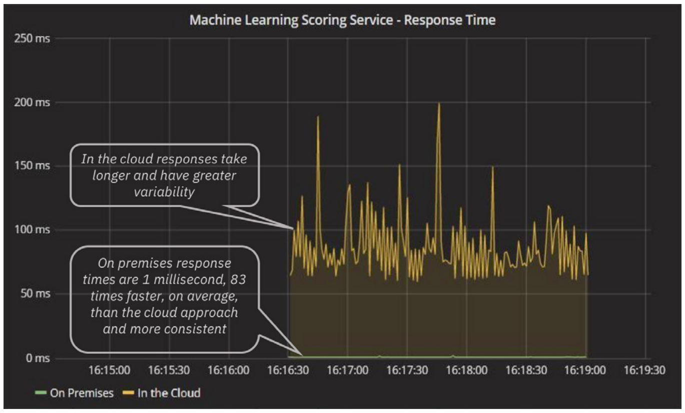

Figure 2: Screen shot showing the inconsistency in response times using the cloud approach

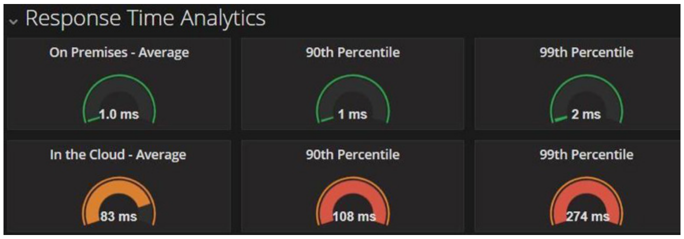

Figure 3: Comparison of average response times on-premises versus in the cloud

Achieve greater overall results by colocating the scoring service with your data

As we've seen by leveraging Machine Learning for z/OS and colocating the machine learning scoring service with the corporate data on the IBM Z platform, organizations can achieve much greater throughput, dramatically better response times and greater confidence in meeting SLAs. When real-time insight matters, the IBM Z platform and IBM z® Analytics deliver the resilient, security-rich, transactional and analytics environment that you need.

For more information

To learn more about the IBM Z platform and IBM z Analytics, please contact your IBM representative or IBM Business Partner, or visit **ibm.com**/z-analytics.

About the authors and contributors

The following people contributed to this thought leadership white paper:

Alex Feinberg
IBM Master Certified IT Specialist, Competitive Project Office

Jonathan Sloan
IBM Analytics Solution Architect, z Analytics

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
August 2018

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary. It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs. THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed, misappropriated or misused or can result in damage to or misuse of your systems, including for use in attacks on others. No IT system or product should be considered completely secure and no single product, service or security measure can be completely effective in preventing improper use or access. IBM systems, products and services are designed to be part of a lawful, comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM DOES NOT WARRANT THAT ANY SYSTEMS, PRODUCTS OR SERVICES ARE IMMUNE FROM, OR WILL MAKE YOUR ENTERPRISE IMMUNE FROM, THE MALICIOUS OR ILLEGAL CONDUCT OF ANY PARTY.