Testing SAS Visual Analytics in a cloud environment using IBM High Performance Services for HPC

Demonstrating IBM's capability to deliver a high-performance cloud environment for analytics using the SAS Visual Analytics package in a test case scenario

Gregg Rohaly, IBM Systems
Ken Gahagan, SAS Institute
Ben Smith, IBM Systems

February 2016

@IBMSystemsISVs

Table of contents

Abstract

This white paper describes an overview of using SAS Visual Analytics in a cloud environment and using IBM High Performance Services for HPC to configure the environment. This paper is written for the technical business manager and IT administrator who are tasked with deploying analytics applications to a cloud topology and would like to read about a proof of concept on this subject. The paper also describes the SAS Visual Analytics performance run and highlights the key performance metrics. It concludes with information about how analytics applications can be run in a cloud environment and how IT staff should consider the use of IBM High Performance Services for HPC to help configure the infrastructure in the goal to achieve optimal performance.

Introduction

Organizations today increasingly make use of solutions such as SAS Visual Analytics to quickly process information and make rapid business decisions. But not every company can afford to have the compute resources to meet the demands of analytic workloads which are intense and variable. Nor do they always have the budget to staff the in-house skills necessary to manage a high-performance infrastructure.

One of the solutions today is to run analytics packages as an on-demand service and take advantage of using a cloud provider. But caution must be used in trying to identify a cloud solution that can deliver a *high-performance environment*. The needs of an analytic system whether under light or heavy load conditions are very demanding and quite different than a traditional transaction-based application running in the cloud.

For these reasons companies have been turning to IBM® and using IBM High Performance Services for HPC. IBM High Performance Services enables quick deployment in the cloud of high-performance computing (HPC) packages such as the SAS Visual Analytics solution. Using IBM High Performance Services for HPC, companies can easily meet additional resource demands without the cost of hiring or managing staff specialist. This minimizes their administrative burden and quickly addresses evolving business needs.

To illustrate these points, a proof of concept was conducted whereby a third-party analytics application, SAS Visual Analytics, was deployed in a cloud environment. To make sure that the requirements were met, the use of IBM High Performance Services for HPC was engaged to help the cloud provider provision and deploy the right private cloud environment using best performance for the minimal cost as one of the criteria constraints.

The conclusion is that it is possible to use a cloud service to offload work that is of analytics nature and achieve performance similar to using a company's local resources. The cloud service can be used for brief periods as demand dictates and special technical skills are not required in-house to maintain these systems, and as a result, the advantages are cost effective.

Assumptions

The reader should be familiar with the operations and management of a data center used in a corporate environment and also with the concept and benefits of deploying workloads to a cloud environment. Additionally, the following topics can help in understanding the deployment of cloud services and conducting the tests. For further reading, refer to the “Resources” section where links to web pages on these and other topics are provided.

SAS Visual Analytics

SAS Visual Analytics Explorer 7.2 is an easy-to-use, web-based product that uses SAS high-performance analytic technologies. SAS Visual Analytics empowers organizations to explore huge volumes of data very quickly to identify patterns, trends, and opportunities for further analysis.

IBM High Performance Services for HPC

IBM High Performance Services for HPC is used to configure and provision high-performance computing applications. When provisioning for a high performance application and its environment, it is important to have access to high-performance computing experts to eliminate the skills barrier for using clustered resources. By working with IBM High Performance Services for HPC, customers can specify dedicated bare-metal servers and high-speed network interconnects, such as InfiniBand® or 40 Gb Ethernet, for applications that require the horsepower of a parallel computing environment. Also, it is possible to specify the location of the data center where it might be required to meet data regulations.

Architecture overview

Note: The architecture described in this paper is just one example of an environment suitable for this test scenario. Other configurations can be designed and used to achieve a similar environment and might yield similar test results.

The architecture design was performed in cooperation with several groups. The first group includes members from the SAS and IBM test team who defined the architecture and hardware requirements. Next, the IBM High Performance Services for HPC team that worked to understand the requirements and translate them to the SoftLayer® cloud team that provisioned the hardware. The SAS and IBM test teams worked to help the other members understand the needs of the application environment. The IBM High Performance Services team took those needs and mapped them into a cloud service designed to provide an optimal infrastructure for the test. The SoftLayer team allocated resources in the cloud to create a suitable platform and infrastructure environment.

From this information, the following requirements are derived for the configuration.

System requirements for running SAS Analytics:

- Design a minimal configuration to achieve optimal performance at an economic cost
- Use an x86 server architecture with the following definitions
  - 4 GB memory per core
  - Local disks for the operating system
  - Fast disks for the data
  - Network speed of 10 Gbps

The IBM High Performance Services team reviewed these requirements and assisted in the correct interpretation of the HPC requirements from the SAS team.

Although it appears to be a simple hardware configuration, deploying to a cloud environment can be challenging. Without specific guidance, the resources might be physically scattered across the data center or even localities. For an HPC application, this could result in poor performance and disappointing results. As the complexity grows, the chances of the cloud deployment being sub-optimal also grows. Without a knowledgeable resource to review an HPC deployment, bottlenecks might occur and that is why companies needing HPC resources in the cloud often find it difficult describing how the resources should configured. The purpose of IBM High Performance Services for HPC is to make sure that the best efforts are used in deployments.

For this scenario, the IBM High Performance team ensured that:

- A high-performance environment was provisioned in the cloud
- All resources were colocated at one data center
- Servers met the specifications
- Hardware was configured, networked, and prepared accordingly

From these requirements the following systems were allocated.

SoftLayer cloud service: Hardware system configuration	
Number of servers	Four bare metal servers, each consisting of
Processors	Two-socket Intel® Xeon® E5-2690 v3 at 2.60 GHz (24 cores)
Memory	128 GB, eight 16 GB DDR4 2133 MHz (PC4-2133)
OS disks	OS disks: Two 1 TB (RAID1)
Data disks	Two 600 GB SAS 15K (JBODs)
Network speed	10 Gbps

Table 1: SoftLayer hardware deployment

The system hardware was assigned the following host names.

Device name	Device Type	Location
c001.sas01.ibmcloud.com	Bare metal	Dallas
c002.sas01.ibmcloud.com	Bare metal	Dallas
c003.sas01.ibmcloud.com	Bare metal	Dallas
c004.sas01.ibmcloud.com	Bare metal	Dallas
vyatta.sas01.ibmcloud.com	Gateway member	Dallas

Table 2: HPC cluster host information

The components of the operating system and the shared file system are shown in the following table. Red Hat Enterprise Linux 6.6 was the operating system. The shared file system provides the ability to migrate data quickly between all of the physical resources. The Hadoop Distributed File System (HDFS) was selected.

Operating system	Red Hat Enterprise Linux 6.6
File system	HDFS

Table 3: Operating system and file system

With the hardware configuration designed, deployed, and enabled, the SAS team can proceed to install the software components. To run SAS Visual Analytics, the following software components are used.

Version	SAS 9.4
Analytics Solution 1	SAS Visual Analytics Server 7.2
Analytics Solution 2	SAS LASR Analytic Server
Metadata Server	SAS Metadata Server
Middle Tier Services	SAS Visual Analytics middle tier

Table 4: SAS software

With the software deployed, the cluster is created and organized into the following systems.

Node 1: SAS Visual Analytics root node

- SAS 9.4
- Hadoop name node
- Middle tier (SAS Web Application Server)

Nodes 2 to 4: SAS Visual Analytics data nodes

- Hadoop and SAS Visual Analytics data nodes
- 72 cores in total for processing
- 384 GB RAM for in-memory data

Test description

SAS Visual Analytics uses resources across a pool or cluster of servers. These solutions use large amounts of RAM, processor, and I/O to solve complex business challenges. A multiuser test scenario is created to simulate an activity within a virtualized environment. A combination of shell and HP LoadRunner scripts are used to control and launch the various workloads.

The performance results that follow are provided for a SAS Visual Analytics 7.2 test scenario. The tests are conducted simulating one heavy user and five light users. The data size is a total of 72 GB.

SAS Visual Analytics Explorer 7.2

SAS Visual Analytics Explorer 7.2 is run on a distributed, four-node, 96-core cluster. There is 72 GB of test data.

Computationally intensive visualizations such as correlations for 10 variables at peak load were run.

Visualizations such as box plots, bar charts, line charts, cross-tabulations, and heat maps averages are run.

Setup conditions for the test

The test is designed to generate a moderate to heavy workload on the server. The goal is to demonstrate the processor usage characteristics and server response time to users' ad hoc analytical requests.

In this scenario, there are two types of users: A light user and a heavy user.

A light user is a business analyst who uses the SAS Visual Analytics server to explore company sales and operational data to quickly discover trends.

A heavy user, whose analytics goal is to rapidly reveal opportunities that can improve revenue or operational efficiency.

The scenario is also designed to approximate the types of usage that might occur during a monthly, a quarterly, or an annual reporting cycle with a mixture of users who need summary reports or graphs and analytical users who need quick answers to questions posed by management.

For this test, there are five light users and one heavy user.

Data description and user base

The storage required for the test data is 72 GB.

The data used for the test can be considered as a large table. The table has 262 million rows and 46 columns.

The content of the data is such that:

- There are more than three years of daily detail at a product description level.
- The geography hierarchy includes facility, region, state, and city
- The product hierarchy includes year, month, date, and description
- The time hierarchy includes year, month, and date
- The measures include revenue, expense [including, capital expenditure (CAPEX), material, operational staffing], employee counts, profit, product quality, and unit capacity

User roles

The two types of users defined are light users and heavy users. Light users typically perform less processor-intensive activities, such as generating summarized reports or simple univariate statistics and graphs. Heavy users perform more advanced statistical analysis such as correlations.

Five light users scenario

For this scenario, five concurrent light users perform the following actions:

- Log in to SAS Visual Analytics Explorer.
- Select the data table, report type, and variables to include for their analysis.
  - Report types include: bar charts, line charts, box plots, cross tabulations, and heat maps.
  - There is between 5 to 15 seconds of think time between drag-and-drop actions.
  - After displaying the report, there is 1 to 3 minutes of think time.
- Create a total of three reports following this process.
- After displaying reports, log off and later log in after random delays of 60 to 90 seconds.

One heavy user scenario

One heavy user performs the following actions:

- Log in to the SAS Visual Analytics Explorer.
- Select the data table for analysis.
- Select 10 variables for a correlation analysis.
  - There is from 5 to 15 seconds of think time between drag-and-drop actions.
  - After displaying the report, there is from 1 to 3 minutes of think time.
- Create a total of three reports following this process.
- After displaying reports, log off and later log in after random delays of 60 to 90 seconds.

Running the test

Load Runner is used to drive an hour long, six-user scenario.

Users enter the processing queue at 10-second intervals, so all users are active in 3 minutes. Users are engaged in report design and exploratory data analysis activities.

As a user's processing cycle completes, the session logs off and is replaced by a new user session, maintaining a full level of concurrency.

The test runs for 60 minutes at full concurrency and ramps down in 3 minutes.

Performance results

During the 60-minute scenario, the response times are as follows:

- 9 seconds or less on average for box plots, bar charts, line charts, cross tabulations, and heat maps.
- 7 seconds or less on average for correlations for 10 variables.

Note: It is the opinion of the SAS and IBM test teams that the response times are reasonable and are consistent with the expectations for this test scenario.

Summary

Testing has shown that cloud resources can be configured to run demanding interactive analytics solutions such as SAS Visual Analytics Explorer.

Based on this effort, it has been demonstrated that the use of cloud resources is a viable option for deploying SAS Visual Analytics applications when best practices and resource requirements are considered during planning.

To assist in the planning stages, the use of IBM High Performance Services for HPC has shown to provide the benefits of a best efforts configuration for execution. This gives organizations an expert service capability that can help to make the best use of resources available.

Resources

The following websites provide useful references to supplement the information contained in this paper:

- IBM Systems on PartnerWorld
  ibm.com/partnerworld/systems

- IBM Redbooks
  ibm.com/redbooks

- IBM Publications Center
  www.elink.ibmlink.ibm.com/public/applications/publications/cgibin/pbi.cgi?CTY=US

- SAS home page
  www.sas.com

- SAS Visual Analytics Explorer 7.2 Users Guide
  support.sas.com.documentation

About the authors

Gregg Rohaly is a former solutions architect and is currently a business development executive in the IBM Systems Group. He has more than 10 years of experience working with IBM Systems. You can reach Gregg at grohaly@us.ibm.com.

Ken Gahagan is a director of research and development at SAS. You can reach Ken at ken.gahagan@sas.com.

Ben Smith is a solutions architect in the IBM Systems Group. He has more than 10 years of experience working with the IBM Systems team. You can reach Ben at smithbe1@us.ibm.com.

@IBMSystemsISVs

Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

INFINIBAND, InfiniBand Trade Association and the INFINIBAND design marks are trademarks and/or service marks of the INFINIBAND Trade Association.

Automagic™, CDNLayer®, Challenging But Not Overwhelming™, CloudLayer®, Flex Images®, KnowledgeLayer®, RescueLayer®, SecurityLayer®, SoftLayer®, SoftLayer® device, StorageLayer®, and The Planet® are trademarks or registered trademarks of SoftLayer, Inc., an IBM Company.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local IBM office or IBM authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in IBM product announcements. The information is presented here to communicate IBM's current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.