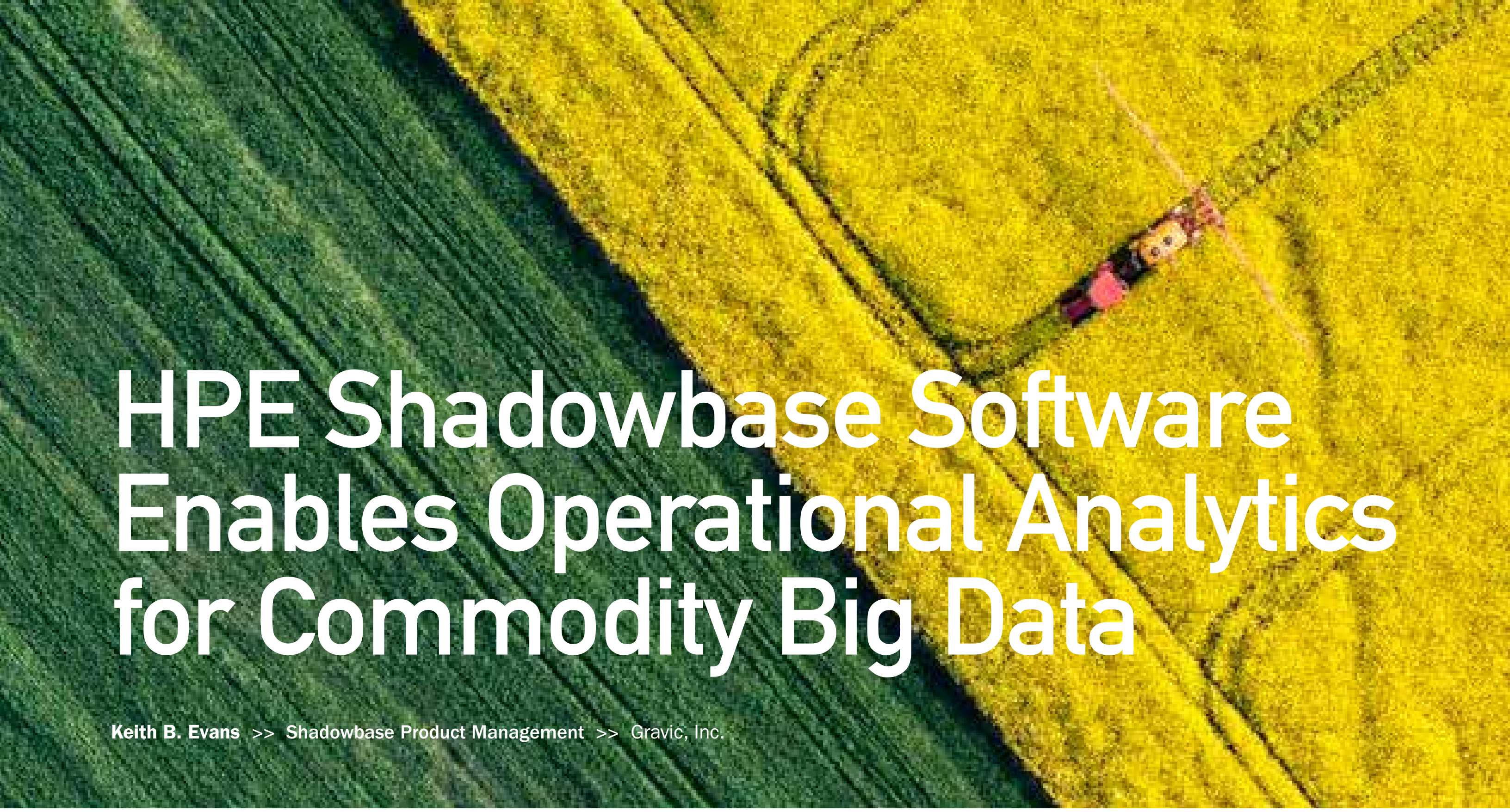


HPE Shadowbase Software Enables Operational Analytics for Commodity Big Data

Keith B. Evans >> **Shadowbase Product Management** >> Gravic, Inc.

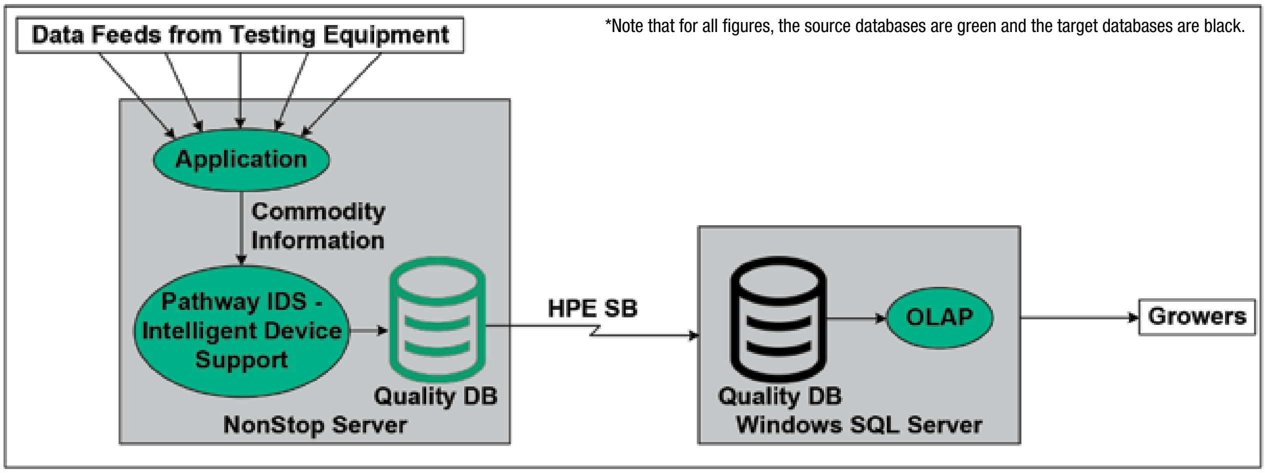
Introduction

The growers of a major U.S. commodity deliver about eight billion pounds of produce per year to consumers. The produce is cultivated at thousands of independent farms throughout the country, and samples used for quality analysis and control are delivered to one of ten regional classification centers run by a large U.S. government agency.

Major commodity producers are shifting towards *precision agriculture* (also known as *satellite agriculture*), which takes the guesswork out of growing crops, shifting production from an art to a science. Precision agriculture is achieved through specialized technology including soil sensors, robotic drones, mobile apps, cloud computing and satellites, and leverages real-time data on the status of the crops, soil and air quality, weather conditions, etc. Predictive analytics software uses this data to inform the producers about such variables as suggested water intake, crop rotation, and harvesting times.[1]

The analyses benefit the growers as a means to improve their products via changes in produce quality, soil moisture content, manufacturing upgrades, etc. The analyses are also used for pricing the commodity product on the open market. However, the analyses can be time-consuming, and the testing equipment can sometimes provide erroneous results as it drifts out of calibration. By the time results are available and are manually reviewed, much of the product has already been distributed to the marketplace; and some of it may have been incorrectly classified.

In early 2016, the agency undertook a major project to create a system that allows quality-control procedures to be performed in near real-time. The new system also provides for aggregation and historical analysis of the quality control information from all ten classification centers. The system's applications employ Online Analytical Processing (OLAP) utilities. Major goals of the project were to distribute data using the OLAP tools in a shorter time frame, to provide immediate notification if any of the testing



*Note that for all figures, the source databases are green and the target databases are black.

Figure 1 – A Regional Classifying Center

[1] For more information, please visit: http://bit.ly/2tkYVJa.

[2] A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed.

Power Outages

The information on commodity quality is also sent by the CS to the futures market, where the crop is bought and sold. The spot price for the commodity on the futures market is set every day at 2 PM, and is partially determined by the quality control information. If a regional classification site loses power from Mother Nature (as the agency often faces) or otherwise goes offline, under the current configuration, samples are sent to another site for processing.

When power returns or the systems are brought online again, the systems are restarted and the Shadowbase software picks up right where it left off. For some systems (not configured with HPE Shadowbase), manual management intervention is required to restart them, creating a time delay, and minor inconvenience.

The central office is the only location with true disaster recovery enabled. However, the agency is considering adding disaster recovery to other systems in order to dramatically increase uptime through a more resilient, zero downtime infrastructure.

equipment became uncalibrated, and to improve the delivery and richness of feedback to growers eager to provide the highest possible product quality.

The HPE Shadowbase data replication engine plays a major role in this system since it automatically moves data from the testing equipment to the multiple OLAP quality analysis applications. Testing results can be immediately validated because the OLAP enabled quality measures occur at the same time. HPE Shadowbase software also creates an up-to-*date data lake*[2] of the classifying centers' output for historical analytical processing.

A Big Data Problem

As the produce is harvested, it is bundled into millions of 500-pound bales, and a sample read from both sides of each bale (with its own unique identifier) is sent to one of the agency's regional testing centers. The agency performs its tests under strictly controlled environmental conditions. The temperature is held at 70° Fahrenheit, and the humidity is fixed at 68%. Several quality measurements are made on each sample, including uniformity, color, moisture content, particle content, purity, and tensile strength. Consequently, over a 100 million measurements are made on the commodity throughout a single growing season.

Historically, the measurement data was printed on green-bar computer paper, and required an analyst to review the results. In addition to the sheer enormity of the printed output, small anomalies were hard to identify and isolate, especially in real or near-real time.

The lengthy time frame between samples being delivered to the agency and the analyses results being distributed to growers is due to the enormity of the processing of the samples and the sheer number of measurements. This delay leads to a significant problem if a manual review of the quality control data – where area directors would have access to commodity samples from each site – shows that some of the commodity has been improperly classified after much of the product has already been delivered to the marketplace.

In addition to growers' eagerness to receive faster, more informative evaluations of their crops, the merchants to whom growers sell their produce advocate near real-time quality analyses as well. For the merchants, price-setting is based on product quality and availability. The quicker the quality and volume can be determined, the faster an equitable price can be set for sale on the open market.

It was clear to the agency that a solution was needed to permit near real-time analysis of the immense number of measurements made during every testing procedure, and to return results rich in useful information to the growers and merchants in a timely manner. The answer was the implementation of a large, distributed OLAP system, which could distill the data to a manageable size, with the ability to quickly identify any aberrations as the testing results flowed through the system. The agency selected the HPE Shadowbase data replication engine to move data from one processing step to another in the distributed system, thus integrating several separate applications into a cohesive whole. This case study explores the use of data replication, application integration, and big data analytics to unlock the value of enormous amounts of operational data.

Analyses in the Regional Classification Centers

The first step in data distillation takes place in the regional classification centers. Each center is equipped with an HPE NonStop system and a Windows SQL Server environment, as shown in Figure 1. The centers use a variety of testing devices to measure the characteristics of each sample. The data from a center's testing devices is sent to its supporting NonStop system.

Next, the NonStop system processes the data via a local application running in a Pathway environment. The data is entered into a SQL/MP Quality Database by Pathway Intelligent Device Support (IDS). From there, it is replicated by the Shadowbase data replication engine (HPE SB) to a Windows SQL server environment for further analysis. The Windows system runs an OLAP application that analyzes the quality-control data for the region. The results of these analyses occur in near real-time and will immediately alarm if any results are out of band from the expected ranges, or if the equipment fails an internal validation test. The results are also returned to the growers via a customized SQL Server-generated view, for guidance in making crop improvements, and to the markets for immediate pricing input. All results indicating there is a problem with a particular batch are flagged for further, immediate review by the agency's team of analysts.

Throughout the agency, thirteen pairs of NonStop/Windows systems are deployed (a single pair's configuration is shown in Figure 1). One pair is located at each of the ten regional sites; one of these sites is collocated with the Central Site (CS), described below. Three pairs are used by the CS: one pair aggregates the data from all the regional sites; another serves as the QA system, and the third is the development system. A disaster recovery system is located about 500 miles away, and can be placed into service should one of the regional systems go down.

TIBCO Spotfire (Figure 2) is a data visualization and analytics software product that runs on an HPE ProLiant server. The agency uses Spotfire to monitor and manage the systems in the regional

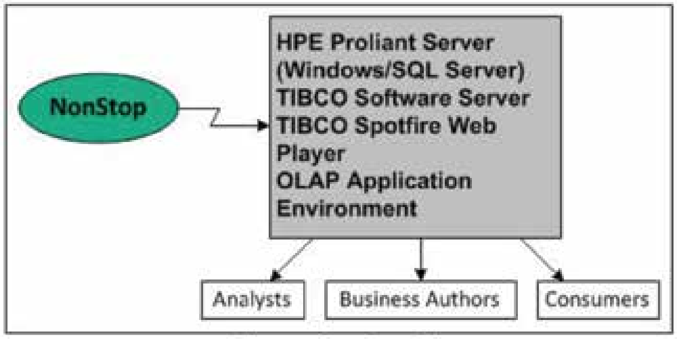

NonStop
HPE Proliant Server (Windows/SQL Server) TIBCO Software Server TIBCO Spotfire Web Player OLAP Application Environment
Analysts
Business Authors
Consumers

Figure 2 – Spotfire

[3] According to Investopedia, "Futures contracts are made in attempt by producers and suppliers of commodities to avoid market volatility." Read more at: http://www.investopedia.com/terms/f/futuresmarket.asp.

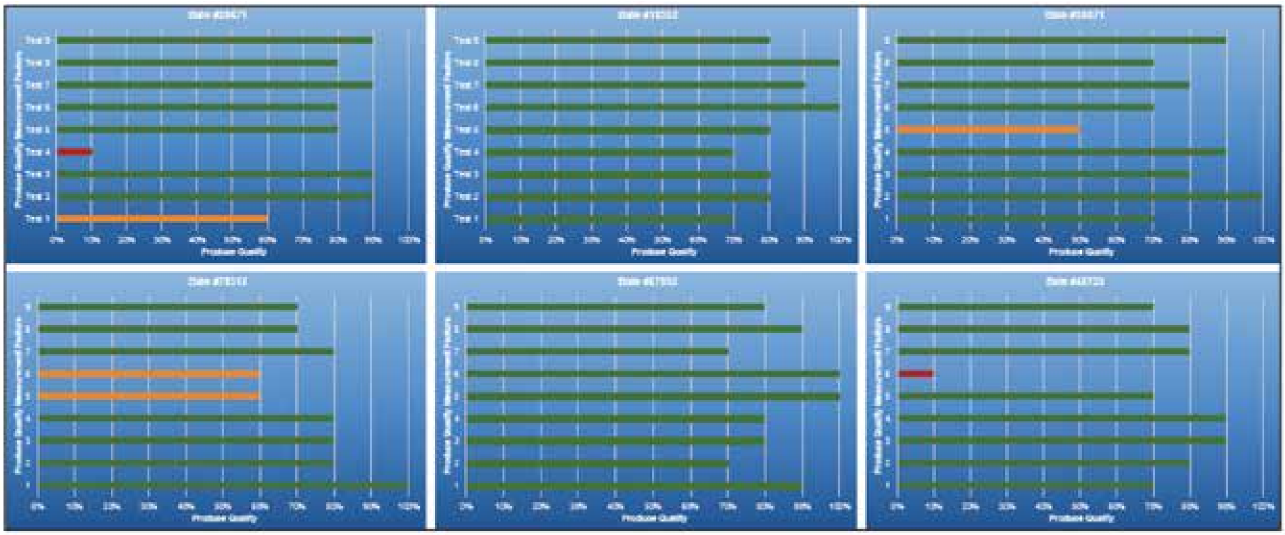


Figure 3 – Produce Quality Dashboard

offices in order to provide a visual dashboard showing the status of the various systems. Spotfire highlights testing results that return out-of-band values. The OLAP software allows the users to point and click, honing in on the raw data for review. Spotfire ensures all systems are working properly and that accurate quality control results are returned to the growers in real-time and before the commodity is shipped.

Spotfire defines three classes of users – analysts, business authors, and consumers – each with a specific role. One problem faced by the agency[3] is that the testing equipment drifts out of calibration over time. Analysts ensure the testing equipment is properly calibrated and that the systems being monitored by Spotfire are working properly. Spotfire indicates the operational status of each system via a series of colors: green means the system is operating properly; yellow means that the system is operating, but it may not be producing accurate results due to improper calibration of the testing equipment; and red means that the system is down, for example, not responding (Figure 3). As a consequence, the health and proper calibration of the testing equipment is continually monitored, and can be recalibrated immediately if it starts to drift or otherwise fails.

Business authors can create and modify the rules that Spotfire supports for analyzing the data, and they can change the graphical and dashboard displays presented by Spotfire. Consumers are the growers and manufacturers who are the end users of the quality control information (e.g., for improving their product or for determining the price they should pay for any particular lot of product). This system also leverages the consumer demand as market feeds, pricing the commodity on the spot market.

The Central Site (CS)

The quality control information generated by the Quality Databases at the regional offices is consolidated and integrated via Shadowbase data replication to the agency's Consolidated Quality Database (CQD) at the CS, which is located in the heart of the growing region. It is organized in a classic hub-and-spoke architecture (Figure 4). The CQD is a SQL/MP database running on a NonStop system (Figure 5).

The CS also houses a Windows/SQL Server environment for processing the aggregated CQD. Shadowbase replication keeps this database up-to-date as the data changes in the regional classification centers. An OLAP application on the Windows/SQL Server environment processes the consolidated data's quality measures for the entire commodity crop in a manner similar to the regional offices.

The regional offices save their data for five years, while up to fifteen years of consolidated data is saved at the CS. In effect, the CQD is an aggregated data warehouse of the quality control information from all ten Regional Classification Centers over several years, allowing for both tactical near-term analyses and strategic long-term analyses. For instance, the agency can look at trends over time for production quality, quantity, and so forth.

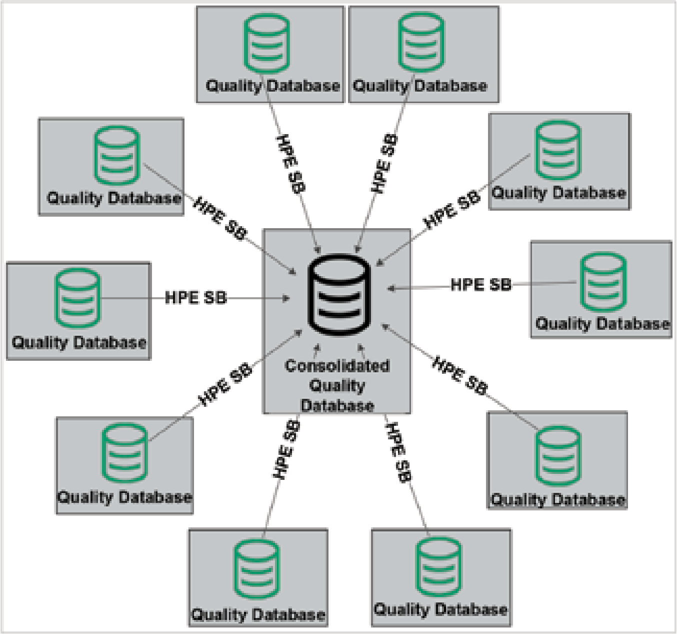


Figure 4 – Consolidated Quality Database

The consolidated information is sold to wholesalers. Typically, they are mills that are looking for specific crop quality parameters to optimize processing procedures and make accurate pricing decisions.

The Many Roles of HPE Shadowbase Data Replication Engine

The commodity application described above is an excellent example of big data analytics using data integration and application integration. Data must be copied between systems and aggregated at the CS. Multiple applications must also interoperate with each other, each extracting quality control

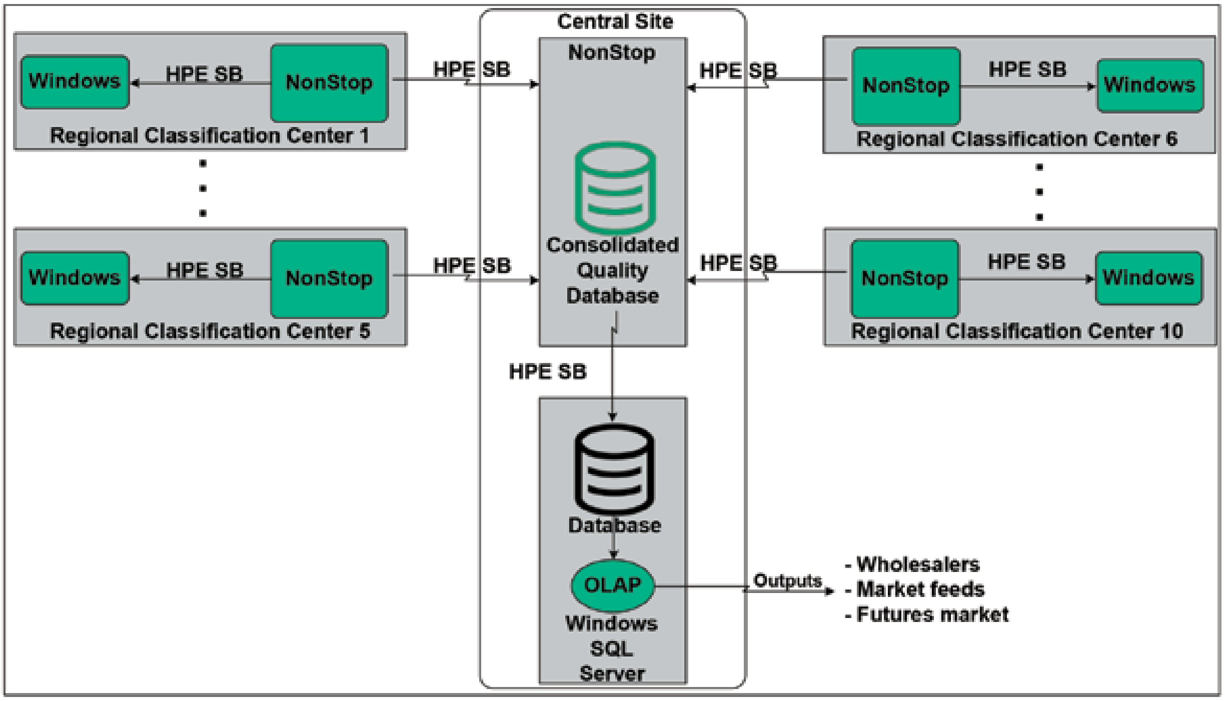


Figure 5 – Central Site (CS)

information from the test data and sending the information to the next application. The Shadowbase data replication engine plays a major role in this solution:

- It replicates test data from the NonStop systems in the agency's regional offices to the Windows SQL Server environments in those same offices for OLAP analysis.
- It replicates the test data to the CQD at the CS.
- It replicates quality control information from the CQD at the CS to a Windows SQL Server environment at the CS for overall OLAP quality analysis, enabling the agency to monitor the crop as a whole.

Conclusion

The commodity quality control system integrates applications to extract meaningful data from a mass of test and product quality data, or Big Data. Extraction is efficient thanks to a series of OLAP processors. HPE Shadowbase software plays a key role in the system by integrating distributed applications through replication of data between the applications. With the right tools, the system can also determine where the commodity with certain quality control parameters (such as tensile strength) was grown and shipped.

The solution has dramatically improved the reporting of the quality metrics for the crop, immediately alarming the users when the parameters are not met. It provides for near real-time tactical reporting of crop metrics to the growers (for production analysis), manufacturers (for manufacturing analysis), and markets (for accurate pricing), and provides historical analysis for strategic purposes.

HPE Shadowbase software (built by Gravic, sold by HPE) is available from HPE and globally sold, supported, and provided service by Hewlett Packard Enterprise. Contact your local HPE Representative for more information.

Mr. Evans earned a BSc (Honors) in Combined Sciences from DeMontfort University, England. He began his professional life as a software engineer at IBM UK Laboratories, developing the CICS application server. He then moved to Digital Equipment Corporation as a pre-sales specialist. In 1988, he emigrated to the U.S. and took a position at Amdahl in Silicon Valley as a software architect, working on transaction processing middleware. In 1992, Mr. Evans joined Tandem and was the lead architect for its open TP application server program (NonStop Tuxedo). After the Tandem mergers, he became a Distinguished Technologist with HP NonStop Enterprise Division (NED) and was involved with the continuing development of middleware application infrastructures. In 2006, he moved into a Product Manager position at NED, responsible for middleware and business continuity software. Mr. Evans joined the Shadowbase Products Group in 2012, working to develop the HPE and Gravic partnership, internal processes, marketing communications, and the Shadowbase product roadmap (in response to business and customer requirements). A particular area of focus is the newly patented Shadowbase synchronous replication technology for zero data loss (ZDL) and data collision avoidance in active/active architectures.