Discovering Meaningful Patterns Using Data Integration

A Gravic, Inc. Article
The Connection, **September/October 2019**
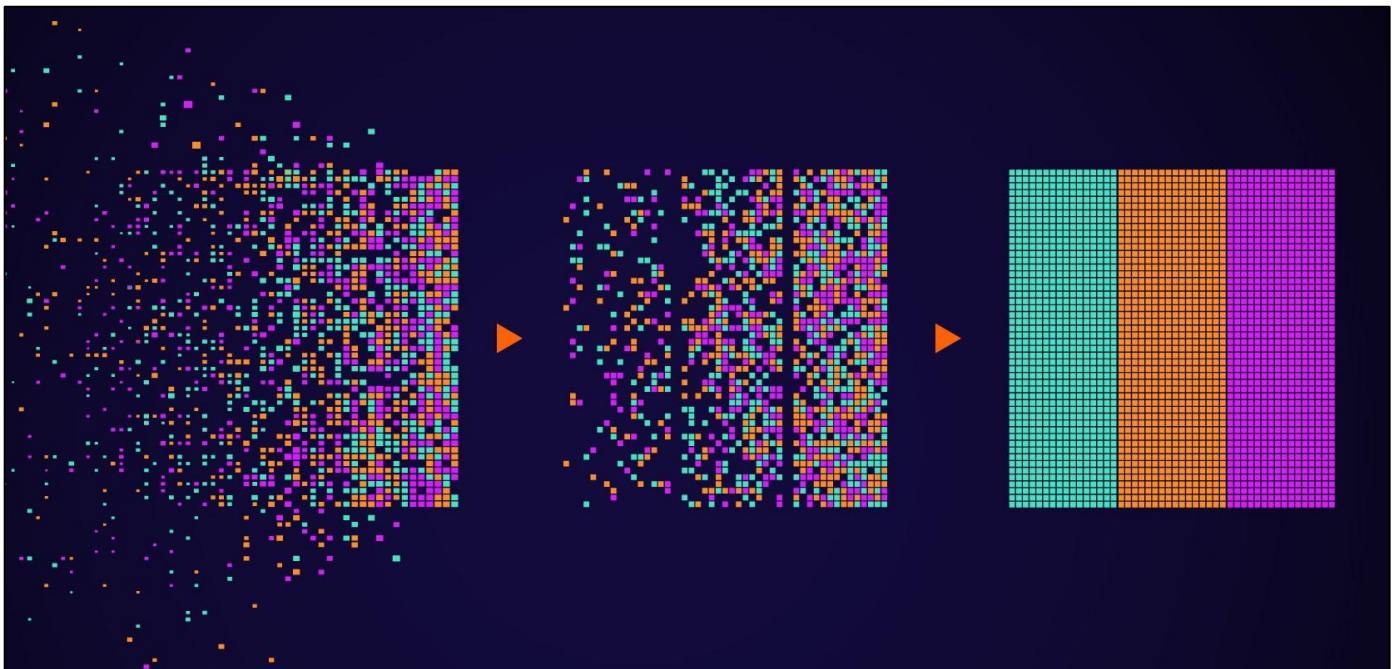**by Keith B. Evans and Paul J. Holenstein**

## Table of Contents

Table of Figures

Introduction

Data is valuable. And the more current the data is, the more valuable it is. However, if this data is trapped or "siloed" in a single system, and not available to enable other real-time business intelligence processes, then its value is limited, and competitive opportunities are missed. To avoid these missed opportunities, companies need to liberate their trapped data in real-time, and make it immediately available to other applications. For these reasons, some of the most valuable data is the data flowing through online transaction processing systems.
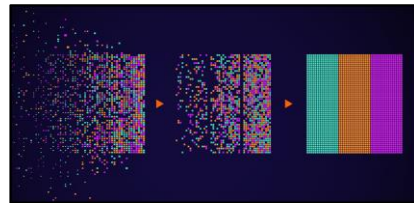
As legacy applications increase over time, siloed data is significantly exacerbated, creating a problem. These legacy applications maintain databases of information, but the data is not exposed for use by other applications. In addition, as companies merge, it becomes necessary to somehow join their disparate databases into a common repository of knowledge for the newly formed corporation, often referred to as creating a *single version of the truth*.

In addition to all these needs, there is now also *big data*. The amount of information being generated each year is exploding at an unprecedented rate. It is estimated that 90% of the world's data was generated in the last two years, and this rate is increasing.[1] By 2020, there will be about 40 trillion gigabytes of data (40 zettabytes).[2] Twitter, Facebook, news articles and stories posted online, the IoT (Internet of Things), YouTube and other videos – they are all accelerating this big data explosion.

Data access is another complicating factor in today's 24x7 environment. Periodic polling or querying for data is slow, system-intensive, and complex. Business needs to react far more quickly than periodically polling for query results. Additionally, running queries against an online database will increase transaction response times, which is typically unacceptable.

"We call data the new currency." – Antonio Neri, HPE CEO and President

---

With an explosion of data at this rate, it is clear that companies who can properly manage their data to extract its value will create new solutions and allow them to quickly surpass their competitors. Within data integration, this management includes feeding a data lake, eliminating data silos, decision support systems (DSS), data warehousing, real-time fraud detection, and fare modeling.

If one application accesses data created by another application in real-time, the applications could share data as if it were local. Furthermore, big data analytics engines require a large network of tens, hundreds, or even thousands of heterogeneous, purpose-built servers, each performing its own portion of the task, yet all of these systems must intercommunicate with each other in real-time. Disparate systems and databases must be integrated with a high-speed, flexible, and reliable data distribution and sharing system, enabling the implementation of powerful Event-Driven Architectures (EDA).

A Data Distribution Backbone

HPE Shadowbase Streams for Data Integration (DI) solves some of these challenges, providing the means to integrate existing applications at the data or event-driven level in order to create new and powerful functionality. It seamlessly copies selected data in real-time from a source database to a target database where it can be used by a target application or data analytics engine. As application changes are made to a source database (by change data capture or "CDC"), they are immediately replicated to a target database to keep it synchronized with the source database. For this reason, data integration is often called data synchronization.

In this model, data replication is transparent to both the source and target applications. On the source side, the application simply makes its changes to the database as it would normally. Shadowbase Streams makes any necessary format changes (transformations) to the data in order to meet the needs of the target application,

---

[1] Source: IBM
[2] Source: EMC

which avoids the need to make any changes to the target application, since it receives the data in exactly the right format. In addition, data can be filtered to eliminate changes that are of no interest to the target application and/or sorted to deliver the changes in the desired order.

Upon delivery, the target application can then make use of this data, enabling the implementation of new event-driven services in real-time to enhance competitiveness, to data mine for new and valuable business insights, to reduce costs or increase revenue, or to satisfy regulatory requirements. Many source or target applications and databases can be integrated in this way, since many heterogeneous platforms and databases are supported. Shadowbase Streams for DI modernizes legacy applications by integrating diverse applications across the enterprise via a reliable, scalable, high-performance data distribution fabric.

Data Integration Case Studies

There is no better way to illustrate this data distribution technology along with the new opportunities and value it brings to an enterprise than by looking at a few real-world production examples.

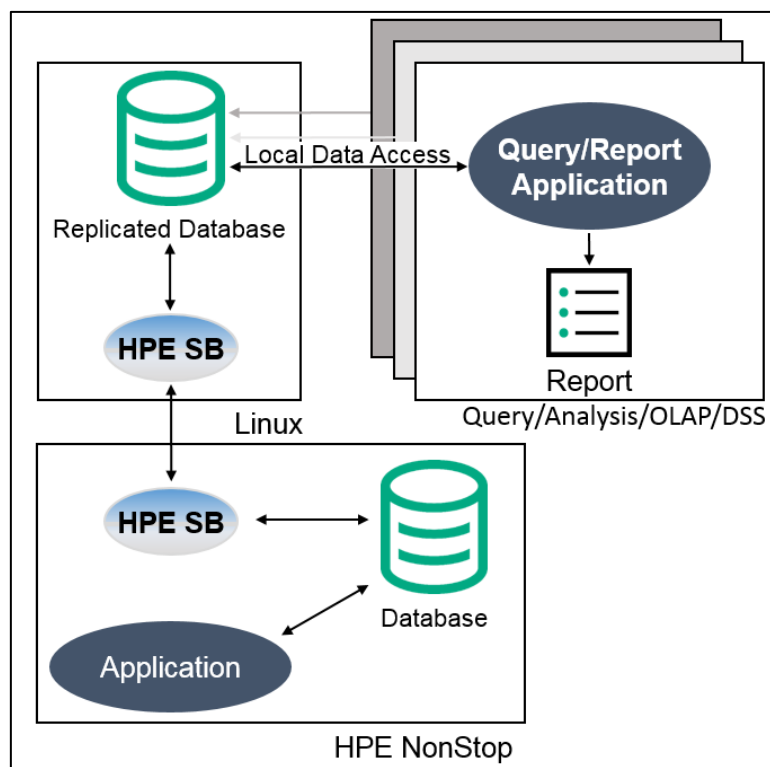*Offloading Query Activity from the Host and Enabling OLAP Processing*

Figure 1 – Offloading Query Activity from the Host and Enabling OLAP

As shown in Figure 1, a large European steel tube manufacturer runs its online shop floor operations on an HPE NonStop Server. To exploit the currency and value of this online data, the manufacturer used to periodically generate reports on Linux Servers using a customized application and connectivity tool that remotely queries the online NonStop Enscribe database and returns the results.

With this original architecture, every time a Linux query/report was run, processing on the NonStop Server was required, and as query activity increased, this workload started to significantly impact online shop floor processing. This impact was compounded due to the high volume of data transformation and cleansing required for converting the NonStop's Enscribe data into a usable format for the reports[3]. Periodically, the company needed to suspend the execution of these reports due to these production impacts. In addition, the remote connectivity architecture was not very robust or scalable, and was also susceptible to network failures, timeouts, and slowdowns when operating at full capacity. Since the data adapter used was nonstandard, the

---

[3] Enscribe is a hierarchical file system native to the NonStop system. As such, it is non-relational and does not have the strong type-casting for data types that most SQL databases enforce. Hence, a lot of transformation and cleansing of Enscribe data is often needed when replicating it into a much more usable relational format that SQL engines can access.

manufacturer could not access the data using standard ODBC[4] and OLAP[5] tools to take advantage of new analytical techniques (such as DSS[6]). Furthermore, the company had a new requirement to share the OLAP analysis with the online NonStop applications in order to optimize shop floor control, which was completely impossible with the original solution. Therefore, a new architecture was required to address these issues and meet the new requirements.

The manufacturer completely rearchitected its Linux-based querying/reporting application. Rather than remotely querying the NonStop Enscribe data each time a report is run, the data is replicated by Shadowbase software in real-time from the Enscribe database to a relational copy of that database hosted on a Linux system. This architectural change only requires sending the data across the network once (when it is changed), instead of every time a query/report is run, significantly decreasing the overhead on the production system.

The raw Enscribe data is non-normalized, and full of arrays, redefines, and data types that do not have a matching SQL data type. As part of the replication process, the non-relational Enscribe data is transformed and cleansed by Shadowbase software into a relational format, and written to an Oracle RAC database. This architectural change only requires cleansing or normalizing the data once (when it is changed), rather than every time a query/report is run.

Since the data is now local to the Linux Servers and presented in standard relational format, it is possible to use standardized SQL data query and analysis tool. In addition, Shadowbase software allows the manufacturer to reverse-replicate the OLAP results and share them with the NonStop applications to better optimize the shop floor manufacturing process.

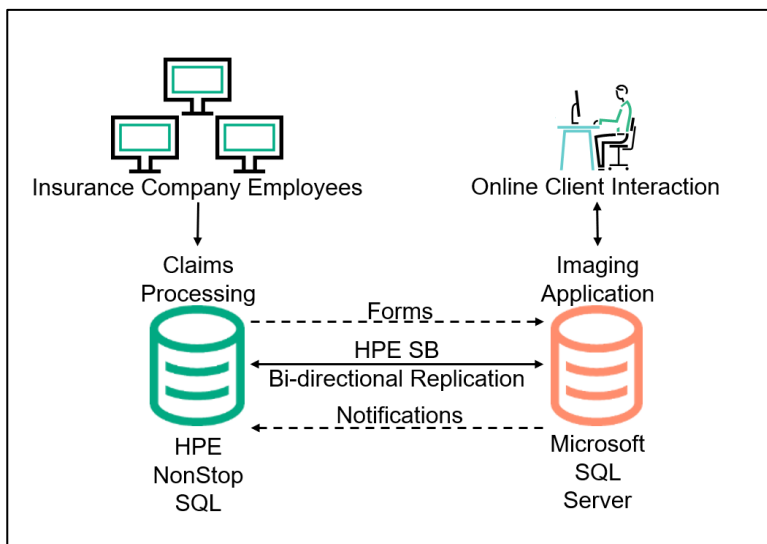### *Optimizing Insurance Claims Processing*

Figure 2 – Optimizing Insurance Claims Processing

As shown in Figure 2, an insurance company provided insurance claims processing for its client companies. It used an HPE NonStop Server to perform processing functions required for each claim and an imaging application to prepare form letters that were the primary interaction with the clients' customers. The imaging application ran on a Windows server and utilized a SQL database. As claims were made, they entered the NonStop claims processing system, then any pertinent information was sent manually to the imaging application on the Windows server. The system created and printed the appropriate letters, which were sent to the customers. The imaging application then returned a completion notification to the claims system, and informed it that the forms were generated.

---

[4] *Open Database Connectivity* (ODBC) is a language-independent, standard database connectivity API containing a library of data access routines. ODBC is very powerful and popular worldwide.

[5] *Online Analytical Processing* (OLAP) is the technology behind many Business Intelligence (BI) applications. OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive "what if" scenario (budget, forecast) planning.

[6] *Decision Support System* (DSS) is a computerized information system used to support decision-making in an organization or a business. A DSS lets users sift through and analyze massive reams of data and compile information that can be used to solve problems and make informed decisions.

This manual process was error prone, duplicates effort, and is time consuming. Also, the company was unable to automatically and quickly meet its clients' needs when checking claim status and verifying claim information. The company wanted a faster and more efficient system that would not only reduce letter production time, but also the number of personnel involved. Shadowbase DI was selected to help automate this process.

In the new application architecture claims are entered into an HPE NonStop application for processing. Then, pertinent forms' information from the claims system is sent to a Windows SQL server imaging application via Shadowbase heterogeneous bi-directional data replication. The imaging application then creates and prints the appropriate forms, which the company mails to the insurance customer, and sends a completion notification via Shadowbase replication back to the claims system, informing it that the forms were generated. This architecture reduces user data entry redundancy and errors as well as increases the speed of claim status updates, enabling much faster processing of claims, all at a reduced cost.

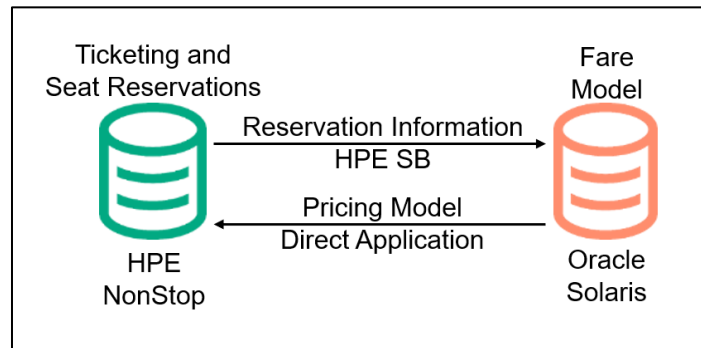### *Rail Occupancy and Fare Modeling*

Figure 3 – Data Integration for Rail Occupancy and Fare Modeling

As shown in Figure 3, European railroad needed to monitor its seat reservation activity in order to create a fare model reflecting the usage of the rail line. The fare modeling was a computationally intensive application running on a Solaris system with an Oracle database.

In the new architecture, an HPE NonStop Server hosts a ticketing and seat reservation application. As tickets are sold, the reservation information is replicated from the NonStop SQL database to an Oracle database on a Solaris system. There, the fare modeling applications track reservation activity and make decisions concerning future fares. The fare modeling application sends back the updated pricing information to the NonStop Server for ticket availability and pricing updates, and then logs consumer demand so that the company can attain price equilibrium for optimum profits.

Both applications and databases are updated in real-time: Shadowbase heterogeneous uni-directional data replication updates the Oracle target with the reservation information, and the updated pricing model information is supplied through a direct application link back to the NonStop. The railroad can now monitor ridership traffic and modify its fare model to maximize profits in real-time.

Prescription Drug Fraud Monitoring
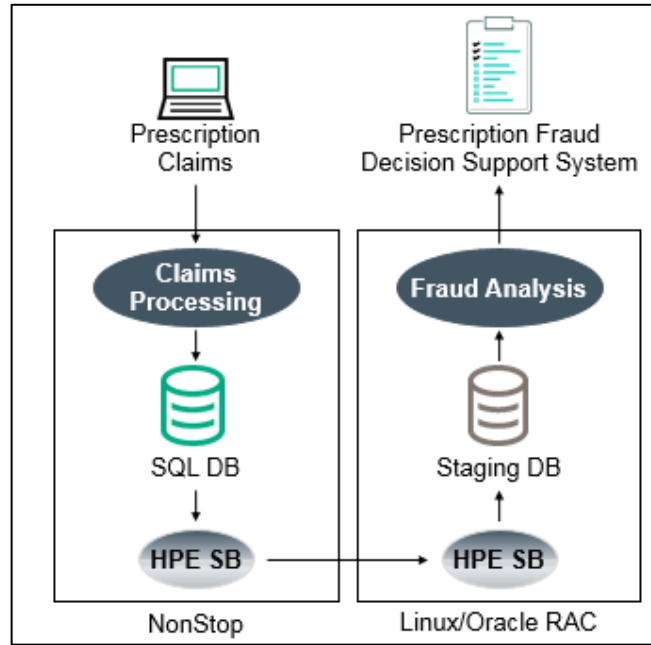
Figure 4 – Prescription Drug Fraud Monitoring

As shown in Figure 4, national healthcare agency processes all prescription insurance claims for the country's healthcare market with a set of applications on HPE NonStop systems. The agency is facing exponential growth in the number of prescriptions due to an aging population along with prescription fraud and abuse. A new integrated system was needed to allow filling valid prescriptions while identifying and flagging fraudulent prescription and reimbursement requests. Due to privacy laws, various jurisdictions maintained data silos, which made it difficult to consolidate data for analysis.

Shadowbase DI was used to integrate the claims processing application with a prescription fraud Decision Support System. This new architecture halts fulfillment of suspicious claims in real-time across all jurisdictions, and provides perpetrator information to law enforcement for monitoring, investigating, and prosecuting fraudulent claims activity. It also stems fraudulent reimbursements to pharmacies, doctors, and patients, saving the healthcare system substantial costs.

Operational Analytics for Commodity Big Data
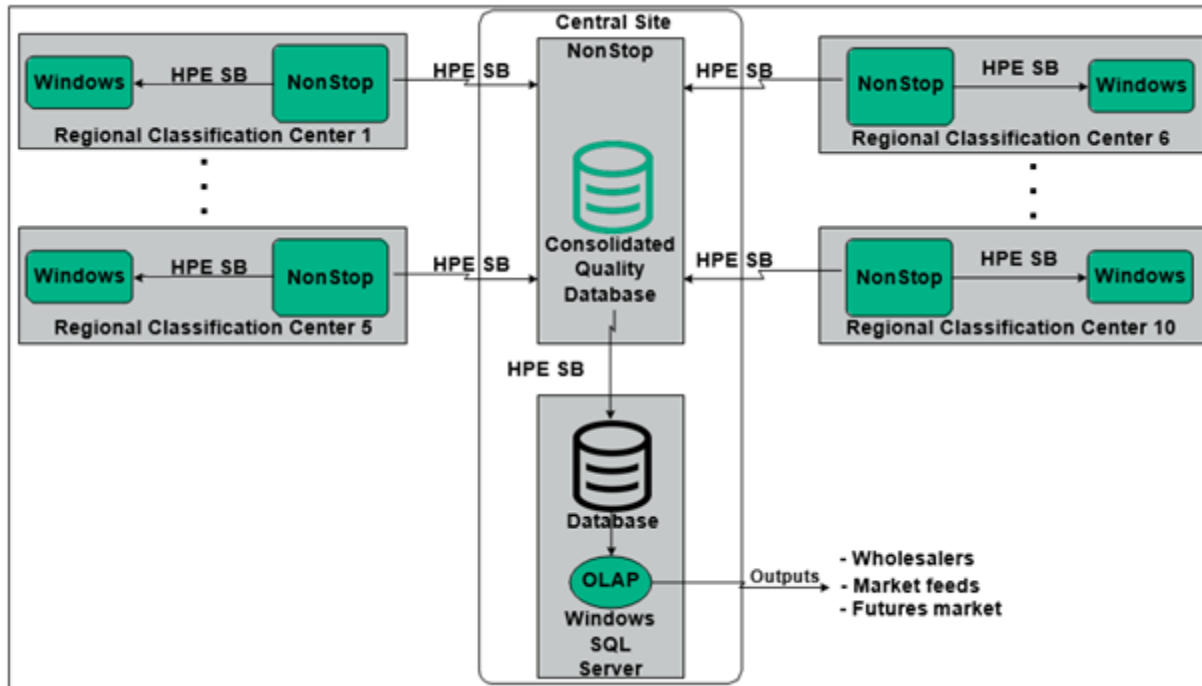
Figure 5 – Operational Analytics for Commodity Big Data

As shown in Figure 5, 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/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 variables such as suggested water intake, crop rotation, and harvesting times.

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 financial spot 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.

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

Shadowbase DI 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 and immediately alarm the agency when testing parameters are not met. Commodity quantity and quality data is distributed in real-time so the produce can be accurately and fairly priced on the financial spot market. Shadowbase software also creates an up-to-date data lake of the classifying centers' output for historical analytical processing.

Summary

HPE Shadowbase Streams for Data Integration uses CDC technology to stream data generated by one application to target database and application environments, eliminating data silos and providing the facilities for integrating existing applications at the data or event-driven level in order to create new and powerful solutions. Applications that once were isolated can now interoperate at event-driven level in real-time. Critical data generated by one application is distributed and acted upon immediately by other applications, enabling the implementation of powerful Event-Driven Architectures (EDA).

International Partner Information

Global

Hewlett Packard Enterprise
6280 America Center Drive
San Jose, CA 95002
USA
Tel: +1.800.607.3567
www.hpe.com

Japan

High Availability Systems Co. Ltd
MS Shibaura Bldg.
4-13-23 Shibaura
Minato-ku, Tokyo 108-0023
Japan
Tel: +81 3 5730 8870
Fax: +81 3 5730 8629
www.ha-sys.co.jp

Gravic, Inc. Contact Information

17 General Warren Blvd.
Malvern, PA 19355-1245
USA
Tel: +1.610.647.6250
Fax: +1.610.647.7958
www.shadowbasesoftware.com
Email Sales: shadowbase@gravic.com
Email Support: sbsupport@gravic.com

Hewlett Packard Enterprise Business Partner Information

Hewlett Packard Enterprise directly sells and supports Shadowbase Solutions under the name **HPE Shadowbase**. For more information, please contact your local HPE account team or visit our website.

Copyright and Trademark Information