opentext™ | Cybersecurity

Position Paper

# Secure & Compliant Test Data Management

## Create, Secure, Share Test Data

## Table of Contents

# Introduction

The year 2020 may well come to be seen as a tipping point in the global adoption of digital transformation. In many organizations, the concept went from theory to practice almost overnight as the pandemic forced a radical shift in working patterns and customer interactions. At the same time, business challenges pressured enterprises to be agile and flexible, to service customers in new markets faster, and to host workloads based on regulatory and geopolitical considerations. These factors are compelling enterprises to adopt the latest technology—including cloud, artificial intelligence (AI), machine learning, mobile apps, automation, IoT, connected devices, cybersecurity, etc.

This has put tremendous pressure on organizations' IT leaders to keep their core business applications up to date in terms of their data (information) and technology. And keeping up with the explosion of data and regulatory compliance mandates from different countries only adds to the weighty responsibility of protecting the data within their boundaries. These rapid technology and regulatory requirements are forcing application development teams to accelerate their application testing to test for the new functionality, integration, and security, etc. To speed up the testing of applications under development and ensure that the right test data are provided, test data is generally created from production data. New attacks and regulatory requirements mandate that organizations identify and remove personal and sensitive content before using test data for its intended purpose.

Recent data breaches and leaks of reputed organizations' sensitive data, as well as fines by countries' regulatory organizations, emphasize the mandate that organizations apply due diligence standards to protecting personal and sensitive data. As shown in the Uber non-production environment case study, insecure software development is a perfect example of how data theft can damage a brand name.

In this paper we provide details on how OpenText™ products and solutions enable organizations to not only generate test data, but also secure it to comply with regulatory requirements and share it across application development and testing teams. You will also find details about the patterns, use cases, and reference architectures used in our Voltage Secure and Compliant Test Data Management (SC TDM) by OpenText™ Solution.


**Figure 1.** Recent Data Breaches

## Challenges in Test Data Management

Application development typically requires maintenance of multiple environments for structured data and code. Development, test, and production are the most common examples. In each of the lower environments, developers and quality assurance staff require sets of test data in order to create new application functionality and execute unit, integration, performance, and system tests. There are many methods for generating test data, including manual generation, functional automation, and data extraction from the production database. Each method has its own benefits and challenges. Our solution concentrates on extracting data from the production database. This method has its own challenges. For example:

- Test data needs to mimic production as closely as possible. Poor data quality or improper obfuscation techniques can lengthen development cycles as developers debug issues related to poor data quality.
- Finding and identifying sensitive and personal data, as well as documenting the remediation of the sensitive and personal data—to provide a defensible position in the event of a breach.
- Keeping test data "fresh." With multiple parallel initiatives occurring in rapidly diminishing DevOps release cycles, "out of date" databases and data structures can cause further delays in development efforts.
- Test data management is typically "disconnected" from DevOps tools and processes. Test data management solutions should not only fit in with your current methodology, but also integrate with existing tools and processes used by the development and testing community.

## Tools Selection for Test Data Management (TDM)

To ensure that their applications behave correctly in the real world and capture a variety of use cases, organizations have historically used real production data—which inevitably includes sensitive personal data—in their testing and quality assurance processes. However, due to the emergence and growth of data privacy laws, organizations can no longer use real production data for testing, development, quality assurance, or education. Therefore, they need the right tools to generate anonymized and protected data and still deliver needed insights.

| Traditional Questions Related to TDM Solutions | OpenText Point of View |
|---|---|
| **Why will test data be created?** | Test data is created to validate new features and functionalities of the applications being developed. It involves using tools and processes for creating a relationally intact, reliable copy or subset of production data, or data very similar to it. |
| **When will the test data be created?** | Based on the functionality or new feature sets of the application, test data will be created during the development cycle of the application. |

*Continued on next page*

| Traditional Questions Related to TDM Solutions | OpenText Point of View |
|---|---|
| **What kind of test data will be created?** | It depends on the application and where it will be used. Generally, it is based on database schema and can be in the form of CSV or structured text files. |
| **Who are the users consuming the test data?** | Generally, it consists of test engineers (functional, integration, system, etc.) and the quality assurance team. |
| **How will the test data be created?** | Based on the requirements, test data will be generated using Manual, Automated Production Copy, and Automated Synthetic data generation. Generally, the process and cost involved in Manual and Synthetic data generation is very high. Automated copy of production data is preferred because it is already in line with the required database schema. However, the requirement of protecting or sanitizing the sensitive data, (per security policies and testing requirements) must be ensured. |
| **Where will the test data be used within the organizational structure?** | The test data will be used by application development teams and cross-functional teams such as testing and quality assurance. |

Current digital transformation needs, and the technology that supports it, have led organizations to evaluate their strategy on TDM solutions that answer these types of questions:

| Questions on TDM Influenced by Digital Transformation | OpenText Point of View |
|---|---|
| **Why do compliance and regulatory requirements need to be considered while generating test data?** | Most countries are working on regulations to secure information related to personal, health, payment, and sensitive business data in use, stored, or while processing. And most of these regulations recommend encrypting or anonymizing data based on the use case. For regulations, such as GDPR and CCPA, anonymizing the sensitive data is recommended while generating test data from production data bases. |
| **What are the risks of semi-compliance or non-compliance with data privacy regulations?** | The evolving enforcement of data privacy regulations mandates that organizations establish the terms of liability and exposure for the sensitive data they handle. The most common approach is to set up a structure of legal actions and financial implications for specific violations. |
| **Why does sensitive data have to be discovered? How can the data discovery and protection in test data creation be automated?** | To protect sensitive data, we must know where they are stored; regulatory requirements also mandate that organizations identify them. OpenText™ products that provide automated search of sensitive data built on technology such as AI and ML and grammars based on regulatory requirements speed up the discovery of sensitive data in all or most data repositories located on premises and in the cloud. |
| **How can existing data be used and monetized effectively in generating test data?** | Different protection techniques can be applied to sensitive data, based on how and where it will be used. When data is encrypted with a protection mechanism, such as Format-Preserving Encryption (FPE), it is easier share the data with the TDM or analytics team because they can simply extract the data from the database. |
| **How can sensitive data be secured to meet approved encryption or protection standards?** | Based on the requirements, test data will be generated using Manual, Automated Production Copy, and Automated Synthetic data generation. Generally, the process and cost involved in Manual and Synthetic data generation is very high. Automated copy of production data is preferred because it is already in line with the required database schema. However, the requirement of protecting or sanitizing the sensitive data, (per security policies and testing requirements) must be ensured. |
| **Where in the geography will test data be used? What kind of compliance and regulatory requirements need to be satisfied?** | For organizations that are geographically distributed (including development and testing teams), they must comply with regulations related to where data is stored and where it is consumed. Fortunately, most of the regulations recommend similar data protection methods (such as encryption or tokenization) for sensitive data. This means the same products/tools can be reused across the organization. OpenText products can be deployed on premises or in the cloud and support integrations with most application platforms. |

In the past, it may have been acceptable to periodically export production data and then import that data into the lower environments. Considering the risk of data breach and regulatory compliance violations, these simple practices no longer satisfy security requirements. Instead, organizations are now pressed to quickly and effectively "protect/obfuscate" sensitive data from a production source as it is moved into a lower environment. Voltage Structured Data Manager by OpenText™ and Voltage SecureData Enterprise by OpenText™ together provide organizations an automated solution to discover sensitive data while extracting test data from production data base, and secure them with Voltage SecureData Enterprise's advanced Format-Preserving Encryption (FPE) by OpenText™ to be compliant with regulatory Requirements.

## Key Benefits and Features of the Secure and Compliant Test Data Management (Voltage SC TDM) Solution

**Key Benefits**

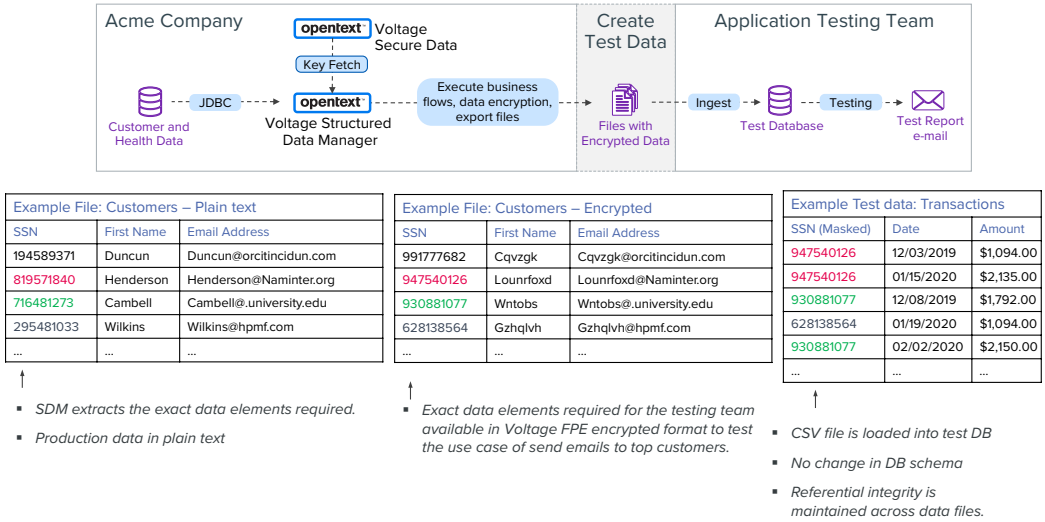
- **Flexible and configurable sensitive data discovery and classification** with prebuilt grammars enable you to reduce your risk by ensuring that you find all sensitive and personal content in the most expedient way possible, thus decreasing your time to remediation.
- **Single cost-effective data privacy** platform to discover, classify, mask, and protect sensitive structured data.
- **Risk scoring** enables you to focus on areas that pose the highest risk, helping you prioritize the reduction of your exposure.
- **Reports of sensitive data locations**, along with their associated risk score for specific targets across the entire enterprise, to speed up your organization's privacy projects.
- **Range of data protection** options to meet complex requirements for test data management, including to comply with various privacy regulations such as GPR, PCI, HIPPA, etc.
- **Secure extraction** of coherent model-based data sets in multiple output formats to migrate into database, CSV, and XML files—for porting into JBDC-supported databases on premises or in the cloud.
- **Reduce risk of data breach**, even when an attacker or malevolent insider can access a database containing sensitive data, the data is encrypted at rest and will be of no value to the attacker, thereby enabling the organization to monetize encrypted-once data with use cases such as test data creation, cloud data analytics, etc.
- **Broad platform support for data-centric protection** use cases can range from traditional transaction-processing systems to cloud-based analytics and SaaS. For example, open systems, cloud, RDBMS UDF, mainframe, mobile, Hadoop, etc.

**Key Features**

- **Automated sensitive data discovery:** The Discovery feature in Voltage Structured Data Manager by OpenText™ supports the discovery, search, and classification of sensitive or personal data. It also enables you to define discovery projects that analyze, declare, and manage sensitive data. The discoveries that are based on business needs can be grouped together into projects. The Discovery feature's grammar defines classes, sets, types, and rules that can be applied to projects based on your organization's privacy and compliance requirements. This feature comes with built-in grammar functions in line with GDPR and PCI-DSS.
- **Create SQL shuffling and test data creation**: The Remediation feature in Voltage Structured Data Manager enables you to create a designer project, which consists of generating masking functions, composition functions, and SQL statements. This feature includes the following functions:
  - **Designer project:** This generates a Voltage Structured Data Manager Designer project to visually customize the data archival process.
  - **Generate shuffles dictionary:** This generates the shuffles for identified sensitive data and also creates a dictionary to be used in test data generation. The dictionary consists of columns identified as "sensitive" during discovery phase. Each column will be generated as a separate dictionary.
  - **Generate encryption:** Built-in integration of Voltage SecureData Enterprise in Voltage Structured Data Manager enables it to use the niche data protection methods such as Voltage Format-Preserving Encryption (FPE) by OpenText™ and Voltage Secure Stateless Tokenization (SST) by OpenText™ provided by it.
- **Masking capabilities in SC TDM:** Voltage Structured Data Manager provides a range of powerful masking options to meet complex requirements for test data management:
  - **Random unique masking:** Uses random values from the dictionary and ensures non-repeatability.
  - **Mapped masking:** Uses referential values in the dictionary.
  - **Mapped unique masking:** Ensures that the values are not duplicated in the generated data and that referential values are maintained.
  - **Random masking:** Uses random values from the dictionary.
- **Data Protection Capabilities:** Voltage SecureData Enterprise is built around its proprietary NIST approved FF1 AES **Voltage FPE encryption algorithm**. Voltage SecureData Enterprise protects data by leveraging:
  - **Voltage Format-Preserving Encryption (FPE)**, which preserves the format of the data being encrypted and enables you to:
    - **Create test data without any changes in source or target database schema.**
    - Leverage FPE-encrypted test data, with **minimal to no application changes**, while using it for testing.
  - OpenText's patented **Voltage Secure Stateless Tokenization (SST)**, which provides protection for payment card data.
  - **Voltage Format-Preserving Hash (FPH)** by OpenText™, which provides a flexible approach to data masking.

- **Sub-setting and Referential integrity:** While creating test data, we generally extract only part or a subset of production data. This leads to the constraint of preserving the referential integrity of subset data.

For example, if a healthcare company has 50 million customers that have 500 million test reports and 500 million transaction reports, test data might require having 100,000 customers, each with 1 million test and transaction reports.




Example File: Customers – Plain text

| SSN | First Name | Email Address |
|---|---|---|
| 194589371 | Duncun | Duncun@orcitincidun.com |
| 819571840 | Henderson | Henderson@Naminter.org |
| 716481273 | Cambell | Cambell@.university.edu |
| 295481033 | Wilkins | Wilkins@hpmf.com |
| ... | ... | ... |

- *SDM extracts the exact data elements required.*
- *Production data in plain text*

Example File: Customers – Encrypted

| SSN | First Name | Email Address |
|---|---|---|
| 991777682 | Cqvzgk | Cqvzgk@orcitincidun.com |
| 947540126 | Lounrfoxd | Lounrfoxd@Naminter.org |
| 930881077 | Wntobs | Wntobs@.university.edu |
| 628138564 | Gzhqlvh | Gzhqlvh@hpmf.com |
| ... | ... | ... |

- *Exact data elements required for the testing team available in Voltage FPE encrypted format to test the use case of send emails to top customers.*

Example Test data: Transactions

| SSN (Masked) | Date | Amount |
|---|---|---|
| 947540126 | 12/03/2019 | $1,094.00 |
| 947540126 | 01/15/2020 | $2,135.00 |
| 930881077 | 12/08/2019 | $1,792.00 |
| 628138564 | 01/19/2020 | $1,094.00 |
| 930881077 | 02/02/2020 | $2,150.00 |
| ... | ... | ... |

- *CSV file is loaded into test DB*
- *No change in DB schema*
- *Referential integrity is maintained across data files.*

**Figure 2.** SC TDM Data Sub-setting and Referential Integrity

The diagram above shows how Voltage Structured Data Manager is used, in a simple example:

- "Acme Company" uses the SC TDM Solution to create test data in order to test a new feature in its healthcare application.
- The plain text is the data form of its production database.
- Test data is created in CSV files with Voltage Format-Preserving Encryption (FPE) applied on all fields.
- Notice in the diagram how Voltage Structured Data Manager can mask the value of the SSN while preserving referential integrity. This is crucial for the use case because a random mask or character substitution mask, such as "XXX-XX-XXXX", would break the business process.
- Voltage Structured Data Manager has the capability to automate and manage highly complex data flows with optional masking of data.

# Solution Architecture—Voltage SC TDM

Our Voltage Secure and Compliant Test Data Management (SC TDM) Solution provides automated test data generation in compliance with security regulatory requirements. Its features include data sub-setting and flexible, secure data protection methods from the organization's production databases. Test data can be created in CSV files or in the target database, depending on the requirements. The solution supports out-of-the box data archival to Oracle, DB2, MS, and SQL databases. It also enables test data creation from databases that support JDBC connectivity in CSV files.

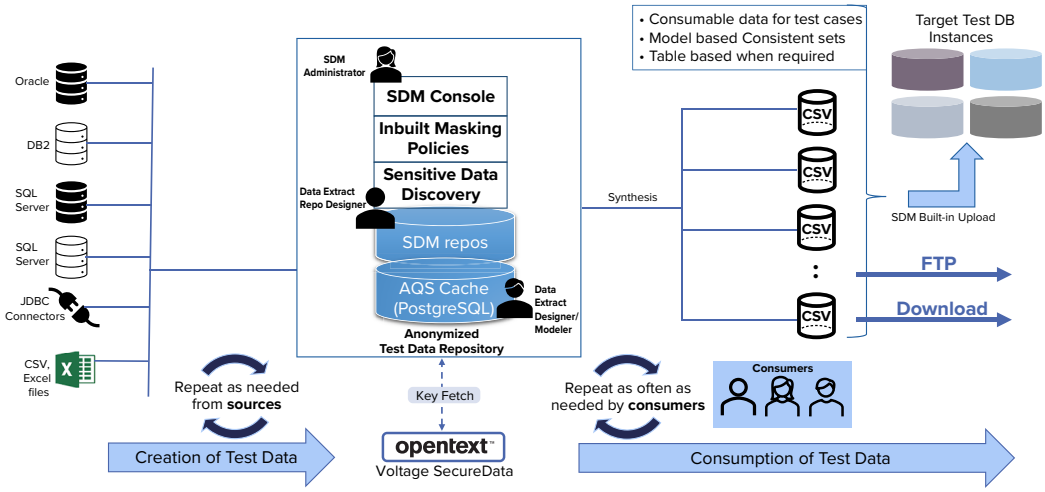



**Figure 3.** Voltage SC TDM Solution Architecture

The architecture of the Voltage SC TDM Solution is shown in the figure above, with details of the components, databases supported, mapping of user personas, and target locations of test data created. Testers can create the test data on premises or in the cloud, as a file or directly into supported bases.

### Solution Components—SC TDM

The Voltage SC TDM Solution has a privacy-based approach that supports a variety of databases. It includes industry-leading data protection methods such as Voltage Format-Preserving Encryption (FPE), Voltage Format--Preserving Hashing (FPH), and Voltage Secure Sessionless Tokenization (SST)—enabling organizations to comply with security and regulatory requirements. The solution comprises Voltage Structured Data Manager and Voltage SecureData Enterprise.

**Voltage Structured Data Manager**

Voltage Structured Data Manager helps to control the growth of mission-critical databases by automating the migration or retirement of data, while preserving its business value and meeting the desired access requirements. It also reduces the total cost of ownership of the application infrastructure, increases business productivity, raises information value, and mitigates the risks associated with tightening compliance requirements. With Voltage Structured Data Manager, data can be relocated to a separate online database for fast, transparent access or to standards-based XML, CSV, or JSON documents for long-term retention—based on retention rules and policies that align with your business.

Voltage Structured Data Manager offers capabilities that address different levels of application complexity, data volumes, and archive access requirements. The components include:

- **Designer:** Provides a visual interface to model data and create business-aligned data migration rules with ease.
- **Data movement:** Ensures that data relocation is performed to meet volume requirements while retaining application integrity.
- **Archive access:** Provides a full range of access capabilities to meet requirements for business operations.
- **Regulatory compliance and legal discovery:** Provides simplified access via a web-based console that doesn't require third-party tools.
- **Job engine:** Automates application archiving tasks with built-in recovery and restart.
- **Management console:** Provides system configuration, job monitoring, job launching, and complete audit trail capabilities. Whether you are running applications on Oracle, Microsoft SQL Server, Sybase, DB2, or open standards JDBC environments, Voltage Structured Data Manager offers the rich set of application archiving capabilities required to control and manage database growth.

**Voltage SecureData Enterprise**

Voltage SecureData Enterprise provides an end-to-end, data-centric approach for enterprise data protection. By leveraging Voltage Format-Preserving Encryption (FPE), Voltage Format Preserving Hash (FPH), Voltage Secure Stateless Tokenization (SST), and Voltage Stateless Key Management, Voltage SecureData Enterprise protects sensitive structured data over its entire lifecycle—from the point at which it is captured and throughout its movement across the extended enterprise, without gaps in security. Voltage SecureData Enterprise "de-identifies" data (rendering it useless to attackers) while maintaining its usability and referential integrity for data processes, applications, and services. Voltage SecureData Enterprise enables the adoption of a continuous data protection model wherever data flows: in analytic platforms and applications, in hybrid multicloud environments, and in native cloud services.

Voltage Structured Data Manager's out of the box integration with Voltage SecureData Enterprise provides data protection in-place or during archival of data for regulatory requirements.

# Voltage SC TDM Process Flow

Voltage SC TDM is a process-based, automated solution to generate test data securely from production databases. The process flow is detailed in the following figure.

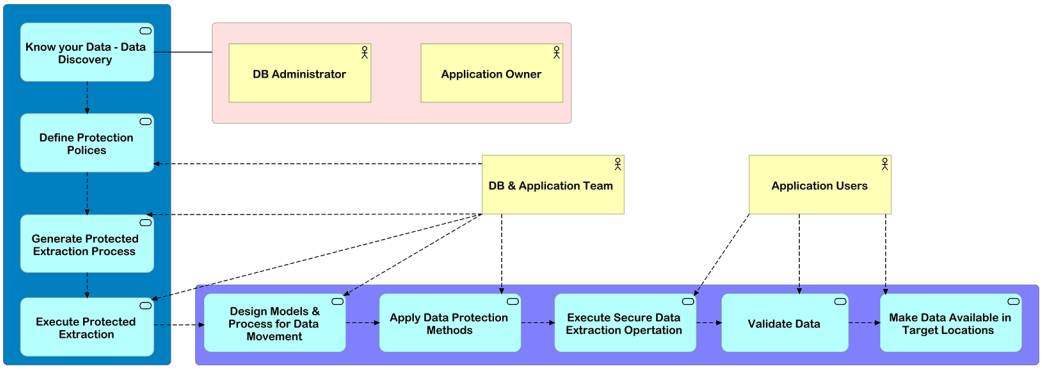



**Figure 4.** SC TDM Process Flow

1. **Know your Data—Data Discovery:** This is the first stage of the secure test data generation process, carried out by the Database Administrator and Application Owners in Voltage SDM WebConsole. Before creating test data from production data stores, organizations must know what sensitive data are present in their data stores with respect to: internal policies, geography-specific security requirements, and regulatory and compliance requirements related to their business. The discovery is an automated process with pre-built and customizable grammars, per the industry's most applicable security compliance regulations (GDPR, CCPA, PCI, etc.).
2. **Define Protection Polices:** Once the sensitive data are present, determine what type of protection mechanisms should be applied in order to protect the data (masking, tokenization, encryption, etc.). This must be done in consultation with the organization's CISO/CTO (to protect the data in accordance with compliance requirements), Application Owners (to ensure that the data produced after protection is usable), Database Administrators (to ensure that the referential integrity is maintained after applying the protection mechanism across data stores, schemas, and tables), and the Testing Team (to determine whether they are able to interpret the generated data in accordance with testing requirements). This is carried out by the Test Manager in Voltage SDM WebConsole.
3. **Generate Protected Extraction Process:** Once the protection policies are applied, the Test Manager generates the automated SQL shuffling functions *(mentioned in the "Key Features" section)*.
4. **Execute Protected Extraction:** After creating the shuffling, the Test Manager creates the Designer project in Voltage SDM WebConsole. (Note: Processes 1-4 are part of the Voltage SDM Discovery module and will be executed from Voltage SDM WebConsole.)

5. **Design Models and Process for Subset Extraction:** This step is carried out in the Voltage SDM Designer interface. In this step, the Test Manager:
   a. Maps columns of sensitive data with identified protection methods.
   b. Specifies the parameter value for the range to specify the data subset size, which will be used by the Test Engineer while executing the Voltage SDM workflow.
   c. Validates the tables in test data creation for referential integrity.
6. **Deploy Secure Subsetting Process:** In this step, the Test Manager finalizes and validates the Designer Cartridges and executes the Business Flow.
7. **OnDemand Test Data Extraction:** In this optional step, the Test Manager creates a new table according to the application's requirements by combining one or more tables or by removing one or more columns in a table and creating test data for that resultant table. The Test Manager also validates the protection methods and referential integrity of data in that table.
8. **Make Data Available:** In this step, the Test Manager deploys the business flow from Voltage SDM Designer to Voltage SDM WebConsole and makes it available to the Test Engineer, who will execute the business flow to create the required amount of test data.

The workflow process is segregated between the Voltage SDM WebConsole and Voltage SDM Designer interfaces.

## Voltage SC TDM Use Cases

This section provides details about how and where the Voltage Secure and Compliant Test Data Management Solution by OpenText™ can be used to provide secure data-centric data migration, securely share data with third parties for analytics, and provide regulatory requirements on data subject requests.

**Use Case 1: Create Test Data from a Production Database**

This use case requires most organizations to extract the data from the current production server, to use in their application development cycle in Test/QA environments. Because data is extracted from the production database, current regulatory requirements mandate that the data is anonymized or masked to protect sensitive data.

Voltage Structured Data Manager reads data from the production database and creates:

- Discoverable sensitive data
- Shuffle-based masking functions
- Data validation for referential integrity
- Data sub-setting
- Encryption of identified sensitive data using Voltage SecureData Enterprise Format-Preserving Encryption (FPE)

Based on the project requirements, Voltage Structured Data Manager applies these to the target archived file and creates test data to be used by test engineers to validate data correctness before its consumption.

**Use Case 2: Data Migration from an On-Premises Database to the Cloud**
Migrating data to the cloud often requires increased examination of the security applied to sensitive data. Using the SC TDM Solution, data is extracted from on-premises data sources, protected with Voltage Format-Preserving Encryption (FPE), and then inserted into cloud data stores such as Dynamo, Mongo, AWS/Azure Data Warehouse, etc. Using Voltage Structured Data Manager and Voltage SecureData Enterprise together, organizations can combine the power of Voltage Structured Data Manager for data extraction and migration with Voltage SecureData Enterprise's advanced Format-Preserving Encryption (FPE). In this use case, data is encrypted (or decrypted) in motion as Voltage Structured Data Manager executes a "business flow." Once it is stored in the cloud, the data can be used as-is or decrypted using other Voltage SecureData Enterprise API integrations available for a range of data warehouse, analytics, and application development environments.

**Use Case 3: Third-Party Data Sharing with Masking**
Sensitive data might need to be shared with third parties for many business purposes, such as analytics, service agreements, and co-marketing. Voltage Structured Data Manager is commonly used to automate the process of extracting data from operational data stores, mask sensitive data elements, and export the resulting data set in the form of a structured data file such as comma-separated values (CSV).

**Use Case 4: Data Minimization**
Newer regulations such as GDPR and CCPA include guidance around data minimization. Data minimization is typically achieved through two objectives:

- Delete data that serves no business purpose.
- Minimize the amount of data stored in operational systems.

## Additional Features of the Voltage SC TDM Solution

- Enables data monetization using secure data migration and archiving.
- Enables data modernization from legacy databases and mainframe databases, such as DB2 and VSAM datastores.
- Automates the process of managing data from old applications.
- Preserves database data for purposes of corporate governance and electronic discovery.
- Scales to meet the needs of the largest and most complex enterprise database applications.
- Provides a long-term data retention solution for production databases.
- Indexes databases to make the data available for searching or query.
- Facilitates business reporting and NetIQ eDiscovery by OpenText™ of structured data.

# Conclusion

Data and application migration to cloud environments, as well as application development in the cloud, require validation that new functional, technical, and business features are aligned with design expectations. In each case, developers and quality assurance staff require sets of test data in order to create new application functionality and execute unit, integration, performance, and system tests. It is common to find that application development and test teams are "cheating" on corporate data privacy rules when it comes to test data management practices. For example, although sensitive data should only be present in production databases with strong access controls, it's often expedient and technically desirable to export production data and then load it into a lower environment for testing purposes. In the past, it might have been acceptable to periodically export production data and then import it into the lower environments. But, considering the risk of data breach and regulatory compliance violations, these practices no longer satisfy security requirements.

Structured data management software can keep these processes in compliance by providing data de-identification as the sensitive data moves from production to QA. For example, Voltage Structured Data Manager and Voltage SecureData Enterprise Format-Preserving Encryption (FPE) not only enable organizations to generate test data from a production database, but also ease the migration of data from on premises to cloud platforms and share the same encrypted data with third parties for data monetization. This enables organizations to encrypt once and use it for multiple use cases.

**Learn More**
For more information on Voltage Structured Data Manager, visit
**www.microfocus.com/en-us/cyberres/data-privacy-protection/structured-data-manager**

For more information on Voltage SecureData Enterprise, visit
**www.microfocus.com/en-us/cyberres/data-privacy-protection/securedata-enterprise**

**Connect with Us**

www.opentext.com

opentext™ | Cybersecurity

OpenText Cybersecurity provides comprehensive security solutions for companies and partners of all sizes. From prevention, detection and response to recovery, investigation and compliance, our unified end-to-end platform helps customers build cyber resilience via a holistic security portfolio. Powered by actionable insights from our real-time and contextual threat intelligence, OpenText Cybersecurity customers benefit from high efficacy products, a compliant experience and simplified security to help manage business risk.

762-000061-004 | O | 01/24 | © 2024 Open Text