Exact Data
Next Generation Systems Testing

Federal Deposit Insurance Corporation
Division of Information Technology

Enterprise Solution for Automated Fully Synthetic Test Data Generation

February 15th, 2016

John Dawson, President
Tel: 585-781-4220
E-mail: john.dawson@exactdata.net
Exact Data, LLC
200 Canal View Blvd., Suite 100
Rochester NY 14623

ITAS II Contract: CORHQ-13-G-0105
(Groups 1 & 2; all subtasks)

Table of Contents

1.0 Introduction

FDIC faces many challenges today in the following areas:

Acquisition Strategy
The ability to quantitatively evaluate a vendor's technical proposal is an on-going problem for every agency. The ability to technically evaluate a solution based on a known test object or database is currently not an option due to the inability to release production databases to the public during the proposal process due to confidentiality issues along with the existence of the applicable database. Management of vendor's technical performance for a solution by Service Level Agreements for speed and error rates is also not possible because appropriate test objects or engineered databases do not exist. Collaborative development is also hindered due to access to appropriate development databases.

Reducing Labor Costs through Automation
Automating time consuming, error prone and expensive manual processes not only improves the time to deploy solutions and quality, but frees up critical resources to work on other mission critical problems and applications.

Compressing IT Development and Implementation Times
Test cycles are a major driver of development and implementation times. Faster testing with concrete test results to feedback into the development and implementation process will significantly speed up these processes. A major time driver for testing is access to the right test databases. Privacy and security issues restrict access and take time to resolve to get access to the right data bases. The database must be audited to determine what is in the database and if appropriate for the specific test case scenario's. Use cases must be established with the data to ensure the specific test cases are included. Today this is primarily a manual process, aided through Extract Load Transform, ETL, technologies and takes a significant amount of time and effort.

Reducing IT Development and Implementation Costs
Reducing overall development and implementation cost allow for strategic investments to better aid the FDIC mission.

Reducing the Cost of Error
Any error will cost time and money to correct and the further downstream they are found from the development process (such as in production environments or at the end users), the more expensive and time consuming they are to correct.

Eliminate Cost and Risk of Managing Private and Confidential Information
Managing private and confidential information is an expensive, labor intensive and costly process with a significant cost associated with mistakes and this information being released to the public.

Cloud Computing
Migrating to the cloud promises to significantly reduce the percentage of information technology budget used for operations and maintenance, enabling agencies to reinvest in, and concentrate on, their core mission objectives. The economic gains of the cloud are so compelling.

The FDIC is currently running approximately 165 different applications and do thousands of patches and releases every year, all with data requirements that are being met through a time consuming manual process requiring data waivers and access to production databases.

2.0 Strategic Alignment with FDIC DIT

We understand that the Federal Deposit Insurance Corporation (FDIC) Division of Information Technology (DIT) released a draft Business Technology Strategic Plan for 2013 – 2017 which emphasizes the strategic imperatives that are needed to provide business value and addresses gaps in either business or information technology capabilities.

The Business Technology Strategic Plan has several key focus areas including Applications Modernization, Strategic Imperatives (Advanced Analytics, Mobility and Electronic Document Management) and Business Agility (Business Process Improvement). Additional detail is provided in **Exhibit 1** below.

Applications Modernization:	Consolidated Applications Modernization Strategy (CAMS) - Address technology obsolescence risk to ensure that business processes are not negatively affected.
Strategic Imperatives:	Advanced Analytics - Harness volumes of data and convert them into actionable insights in order to help drive faster and better decision-making, expedient analyses, predictable outcomes, and optimal operational efficiency
	Mobility - Develop a "mobile first" strategy to support access to high-quality information and applications anywhere, anytime and on any device
	Electronic Document Management - Improve the efficiency and reliability for electronic document processing and workflow automation
Business Agility:	Business Process Improvement - Deliver systems and services in a timely, effective manner to support accomplishing the mission of safeguarding the U.S. financial system

Exhibit 1: FDIC DIT Business and Technology Strategic Plan key focus areas

The proposed solution in this white paper will have a significant positive impact on the focus areas of Applications Modernization, Strategic Imperatives (Advanced Analytics, Mobility and Electronic Document Management) and Business Agility (Business Process Improvement).

3.0 Proposed Enterprise Solution for Automated Fully Synthetic Test Data Generation

An Enterprise implementation of automated fully synthetic test data generation will create the ability to immediately generate virtually unlimited amounts of relevant test data designed for the specific system under test along with metadata descriptor files of the test data.

A major change in the way solutions are acquired can now be implemented through the ability to quantitatively evaluate a vendor's technical proposal. Databases can be manufactured to specifically measure performance of technical requirements. The synthetic databases can be released to the public during the RFP process. The Vendor's technical performance can now be managed through quantifiable Service Level Agreements. Collaborative development is also now enabled between the various industry partners providing solutions.

Automated synthetic data generation eliminates the time consuming, error prone and expensive manual/ETL processes and not only improves the time to deploy solutions and quality, but frees up critical resources to work on other mission critical problems and applications.

Automated synthetic data generation enables faster testing with concrete test results to feedback into the development and implementation process while eliminating the cost, risk and time to manage private and confidential information.

Overall development and implementation cost are significantly reduced through the use of automated synthetic data generation allowing for strategic investments to better aid the FDIC mission.

Synthetic data generation creates large volumes or precise test data which makes the development process more efficient and precise and will reduce the cost time and money to correct and downstream errors.

Fully synthetic test data eliminates the cost and risk of managing private and confidential information.

Migration to the cloud is enabled through the ability to establish test, development and collaborative sand box environments for the vendor community that is fully synthetic and cannot be compromised as this data was never real.

Solution features and benefits to the FDIC include:

Solution Features	Manual Test Data Process	ETL Test Data Process	Automated Fully Synthetic Test Data Generation
Volumes	Very Limited Volumes, Less than 100 Records	Unique Record Volumes are Limited to Original Data Base	Virtually Unlimited, 100M's or 1B's
Time to Market	Months	Months	Weeks versus Months for Initial Data Sets, Hours versus Months for Subsequent Data
Data Integrity and Quality	Poor Quality, Very Difficult to Create Large Realistic Record Sets	Removing Data Elements Destroys the Data Record Internal Consistencies. Ground Truth (content of the data records) is Not Known	Realistic: Contextually Correct (i.e. family groupings), Statistically controlled (i.e. distributions), Longitudinally Correct (i.e. consistent over time). Ground Truth Perfectly Known and Described with Meta Data.
Error Processing	Insufficient Volumes to Measure Error Rates	Use Cases Manually Created with Limited Variation and Not Broadly Interconnected	Data is Engineered for the System Under Test: Interwoven Complex Use Cases and System Answer Files. Enables SLA Agreements.
Privacy and Security		Does Not Eliminate Security and Privacy Risks	Eliminates Security and Privacy Risks, Enables Cloud Implementations and Collaborative Development Environments. Enables Technical RFP Evaluations with Requirements Engineered Test Objects.
Future State		Cannot Create Future State Data, Requires Access to an Existing Data Base	Can Create Any Future or Past Data Base
Cost	Major IT Cost Component	Major IT Cost Component	Orders of Magnitude Less Cost

Exhibit 2. Solution Features and Benefits to FDIC

4.0 Case Study/Experience

The IRS ACA program had originally implemented a process using an ETL solution, Mathematica, to create test data. 6 months into the process, costs were estimated to approach $1M and less than 1M test records had been generated. The quality of the records was very poor, with most of the records unable to pass the ingest process into the new ACA system for processing.

The IRS ACA program then contracted with ExactData™ to provide a 300M record database for use during a performance test in May/June of 2013. Data models were established over a three week period and data samples provided on an iterative basis to the IRS ACA team for validation. Additional data bases of similar scope and size can now be generated and delivered in less than 24 hours. IRS achieved the following comparative benefits by leveraging ExactData.

Positive Results for the FDIC

Reduce the cost of test data creation by 90%. A typical large agency has over 500 full time equivalents involved in this process yielding savings of $50M per year.

Reduce project development timelines by up to 50%. This directly translates to cost savings for that program that will be measured in $100M's per year.

Increase use case coverage and measure error rates in your development environment. Correcting errors in the development process and reducing the cost of escape errors will save your agency $100M a year.

Eliminate the cost and risk of managing confidential and private information in your development environments.

Create collaborative Industry environments where challenge databases are released to Industry to prototype and prove value before you purchase.

Mathetica ETL Soluton	ExactData Solution
• Less than 1M records	• 300M records
• Over 6 months to create	• 3 weeks for the first 300M, subsequent databases with similar characteristics several hours
• Approximately $1M in costs	• 70% less cost
• Poor quality records, unable to ingest most records into the processing system	• Perfect quality of all records with no ingest processing issues
• The original database was form the Center for Medicaid and Medicare Services, CMS and personnel information could be recreated creating security and privacy risks	• Fully synthetic data cannot be re-identified and has no security and privacy risks

Exhibit 3. Qualitative and Quantitative Benefits

The IRS also helped to develop a business case for an Enterprise wide adoption of the technology with the following results contained in Exhibit 4:

Improvement Area	Financial Value
Reduction in Costs over Manual or ETL Process	≈$6M Year One (Reduction in Current 100 FTE's)
Reduction in Costs to Correct a Downstream Error (Corrected in Production versus Development)	≈ $66M Annually (10% Improvement at $14/Error)
Reduction in Costs of Escape Errors (Error Leaves System)	≈ $325M Annually (5% Improvement in Fraudulent Refunds)
Reduced Development and Systems Implementation Times	At Least 30% of Total Information Technology Costs are for Testing, and we can Generally Improve Testing Efficiencies by up to 50%.
Transform the IRS's ability to Make Informed Purchasing Decisions and Quantitatively Manage Service Level Agreement Contracts	Census 2010 Use Case: Over $100M Savings for the Bureau of Census
Compliance with IRS Policy/Eliminate Cost and Risk of Managing Private and Confidential Information. Enabling Cloud and Collaborative Development	

Exhibit 4: Financial Benefits to the IRS for an Enterprise Implementation

5.0 ExactData

ExactData is the world's leading provider of Fully Synthetic Data Manufacturing Services. Current clients include some of the largest US Federal Agencies including the Department of Defense and Internal Revenue Service.

ExactData's correlated, complex "smart" test data is unmatched by any other company in the world. We manufacture our customized test data using a patented, sophisticated rules engine designed for the specifics of your system's requirements for realism, complexity, and scale. Our fully engineered data includes the unique features of longitudinal consistency, internal consistency, consistency across disparate data sets, and perfectly known ground truth. This enables comprehensive quantitative system performance measurement, scoring for error rates, and algorithm testing.

ExactData

We manufacture test artifacts that simulate components of the entire FDIC operation, including associated metadata and paradata, with the attributes of realism, scale, scenario ground truth (data that is generated for each dataset and is precisely known by field) and complex interconnectivity. We make possible high precision measurements of system accuracy and performance, to enable efficient and effective root cause analysis. As system adjustments are made we can adjust in lockstep, providing updated test artifacts virtually on demand.

No one else in the world can do this, as demonstrated by our patented technology, the Dynamic Data Generator™.

For the FDIC, we will enable a new world-class level of system development, testing and iteration at production scale

ExactData generates test data with absolutely no confidentiality or privacy risks. Our unique

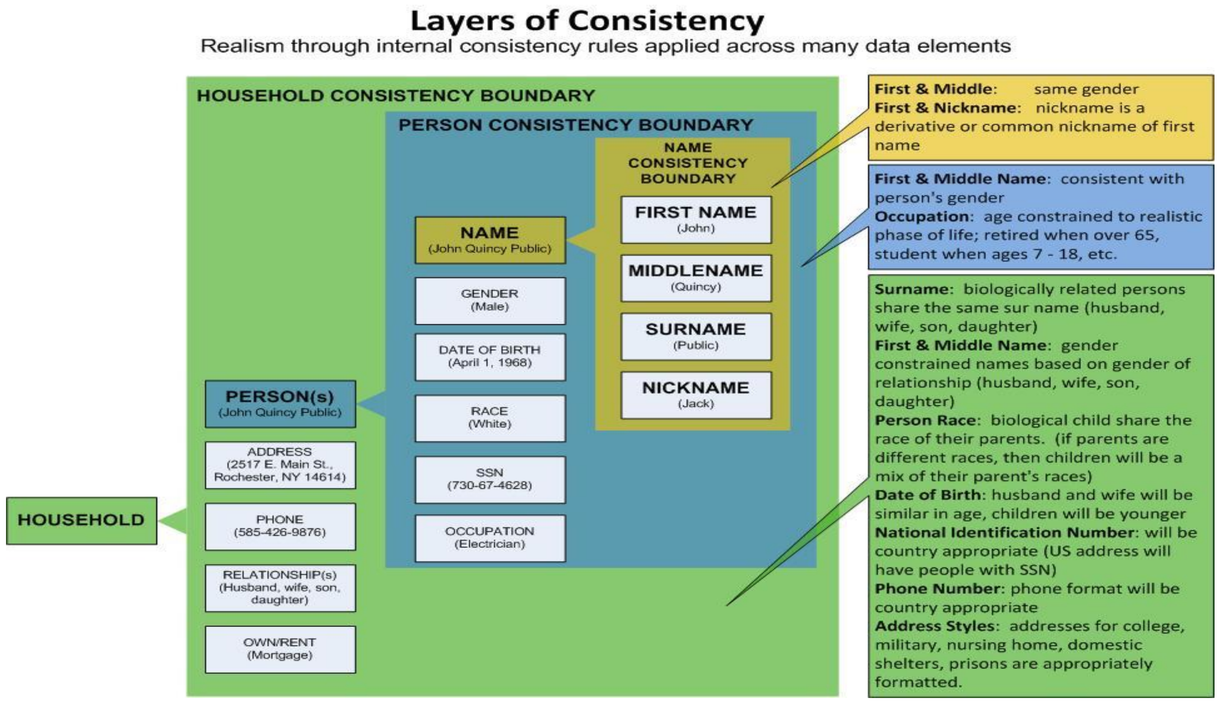
Dynamic Data Generator™ technology doesn't require any access to production data or live records. Using this high-fidelity data, the FDIC can test all system modules, and also test their interfaces as well as the end to end system workflows for seamless and high quality performance. Many other Government Agencies have successfully used ExactData's data for testing: DARPA (for insider threat detection), U.S. Army (DoD Manufactured Artificial medical records), and IRS (Manufactured Artificial data for testing the additional complications of the Affordable Care Act on tax forms). ExactData has supplied synthetic data to commercial enterprises also, such as IBM (Manufactured Artificial data for testing the telephony application in the 2010 Census), and more recently Oracle (Manufactured Artificial medical records for demonstrating their internally developed software).

We do not start with live, actual or production data and alter it but rather manufacture test data using a patented, sophisticated rules engine configured for the specifics of your system test requirements for realism, complexity, and scale. We do NOT de-identify, mask, or otherwise manipulate data. Our data is built from the ground up. It inherently cannot be re-identified because it was never real to start with.

Our approach is to generate a large "universe" of interrelated data and then extract file types and artifacts desired for test. This ensures consistency across all the various data sets that can be derived. For example, we initially utilize complex databases (seed data) of last names, male and female first names, street names, city names, races, etc. chosen probabilistically to produce realistic, but not real households.

Virtually every one of the data characteristics can be altered and are potential for parameterization, with many already parameterized today.

The following chart provides insight into how we maintain data consistency, and for this example at the household level. A Household is just one example of the many "superfields" in our system.


Layers of Consistency
Realism through internal consistency rules applied across many data elements
HOUSEHOLD CONSISTENCY BOUNDARY
PERSON CONSISTENCY BOUNDARY
NAME CONSISTENCY BOUNDARY
FIRST NAME (John)
MIDDLENAME (Quincy)
SURNAME (Public)
NICKNAME (Jack)
NAME (John Quincy Public)
GENDER (Male)
DATE OF BIRTH (April 1, 1968)
RACE (White)
SSN (730-67-4628)
OCCUPATION (Electrician)
PERSON(s) (John Quincy Public)
ADDRESS (2517 E. Main St., Rochester, NY 14614)
PHONE (585-426-9876)
RELATIONSHIP(s) (Husband, wife, son, daughter)
OWN/RENT (Mortgage)
HOUSEHOLD

First & Middle: same gender
First & Nickname: nickname is a derivative or common nickname of first name

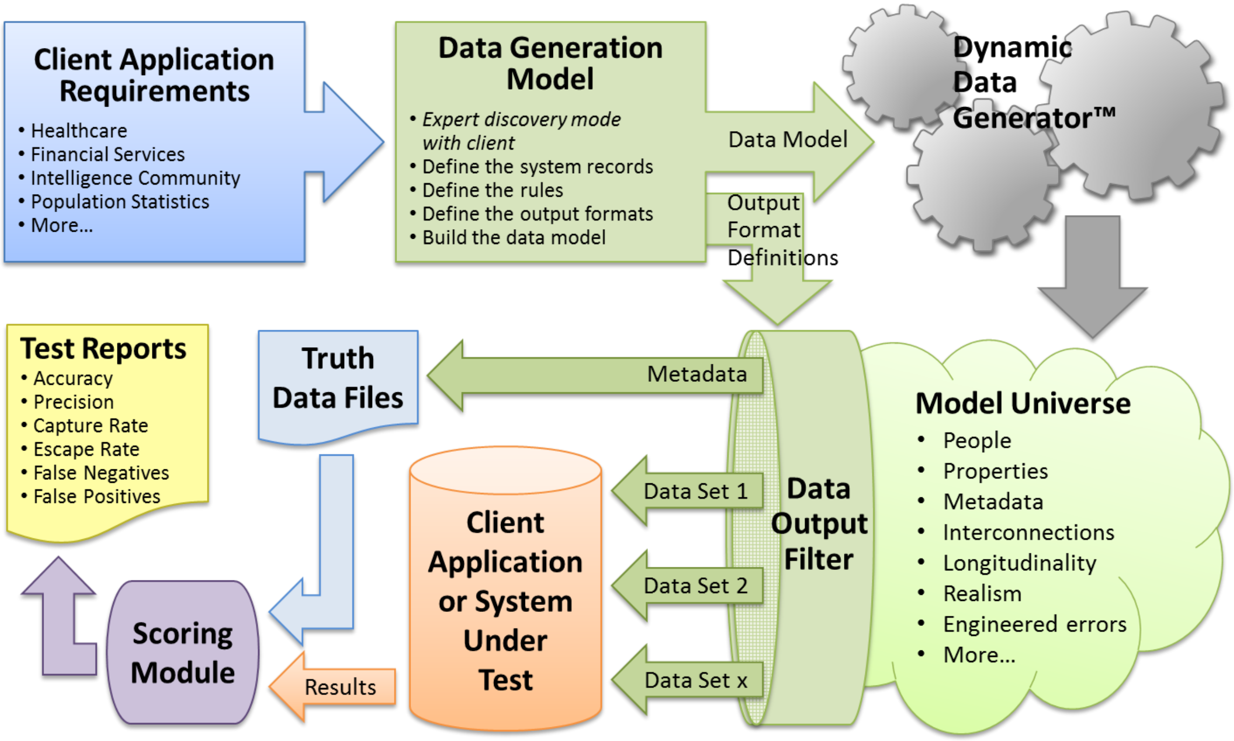
First & Middle Name: consistent with person's gender
Occupation: age constrained to realistic phase of life; retired when over 65, student when ages 7 - 18, etc.

Surname: biologically related persons share the same sur name (husband, wife, son, daughter)
First & Middle Name: gender constrained names based on gender of relationship (husband, wife, son, daughter)
Person Race: biological child share the race of their parents. (if parents are different races, then children will be a mix of their parent's races)
Date of Birth: husband and wife will be similar in age, children will be younger
National Identification Number: will be country appropriate (US address will have people with SSN)
Phone Number: phone format will be country appropriate
Address Styles: addresses for college, military, nursing home, domestic shelters, prisons are appropriately formatted.

At a core level we have a data generation engine. As a data source, we have lists of seed data that the data generation engine pulls from. Our seed data is refreshed on an as needed basis. We then build a template which provides instructions to the data generation engine. Parameters within the template are configurable, literally hundreds of configuration points, including output data format targeted for ingest into system under test. The data generation engine is a tireless bookkeeper that ensures data consistency, internally to each record and longitudinally over multiple records, over even millions of records.

The data can be refreshed (regenerated) as often as desired, virtually on demand (the actual datasets produced are not refreshed themselves, but rather new sets would be generated to meet immediate needs). You could, for example, produce a new data run, with updated date ranges. You could introduce an enlarged test universe, with, say, Manufactured Artificial Social Security and/or IRS data. A great strength of our technology is the capability to repeatedly refine the data model to meet new requirements.

ExactData™ Artificial Test Data Process



Client Application Requirements
- Healthcare
- Financial Services
- Intelligence Community
- Population Statistics
- More…

Data Generation Model
- *Expert discovery mode with client*
- Define the system records
- Define the rules
- Define the output formats
- Build the data model

Data Model

Output Format Definitions

Dynamic Data Generator™

Model Universe
- People
- Properties
- Metadata
- Interconnections
- Longitudinality
- Realism
- Engineered errors
- More…

Data Output Filter

Metadata

Data Set 1

Data Set 2

Data Set x

Truth Data Files

Client Application or System Under Test

Results

Scoring Module

Test Reports
- Accuracy
- Precision
- Capture Rate
- Escape Rate
- False Negatives
- False Positives

12

6.0 Pilot Project Proposal: FDIC Longitudinal Financial Institutions Manufactured Synthetic Data Pilot

Proposal Concept: Pilot project to create a haystack of synthetic financial institutions with longitudinal attributes of mergers, growth, data migration and other changes since 1933 or institution inception date.

Foundation Data:

- FDIC Regional & Field Offices: FDIC to decide synthetic or actual field offices.
- FDIC Personnel: synthetic

Application Data:

- Financial Institutions
- Financial Institution History
- Financial Institution Personnel
- Business Rules: FDIC chooses 50 - 75 business rules for the pilot that will provide evidence of synthetic data viability