Enterprise Challenges of Test Data

Size, Change, Complexity, Disparity, and Privacy

RBCS

TIME TESTED.
TESTING IMPROVED.

www.RBCS-US.com

Enterprise Challenges of Test Data

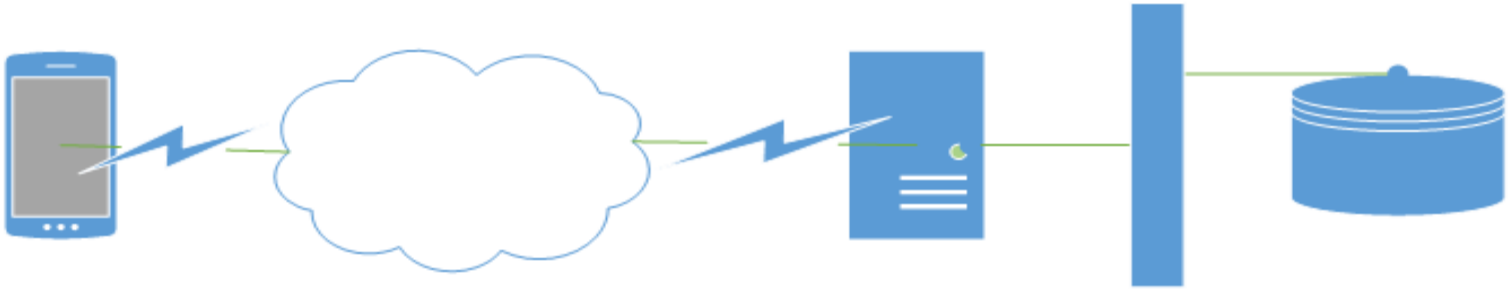
- For simple applications, representative test data can be relatively easy
- What if you are testing enterprise-scale applications?
- In enterprise data centers, dozens or hundreds of applications co-exist
- These applications are of various sizes, complexities, and criticalities
- They operate on various data repositories, in some cases shared data repositories
- In some cases, disparate data repositories hold related data
- Let's examine the test data options, and the challenges that arise, at the enterprise scale

Options for Test Data

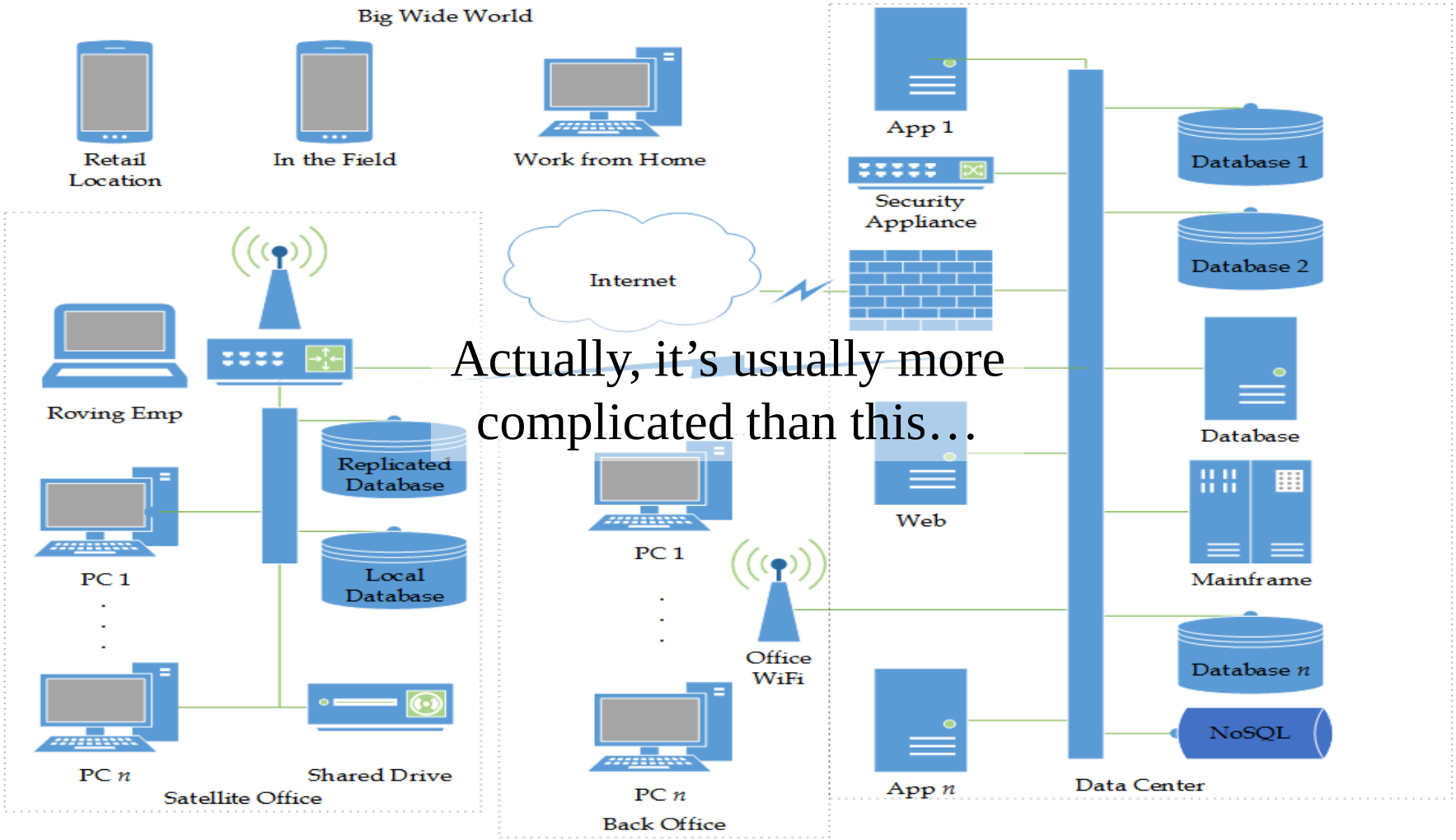
- Hand generated: Tester creates all test data by hand
- Tool generated: Generic (e.g., Excel) or specific tools, open-source or commercial, used to generate test data according to tester-specified pattern
- Production: all or a subset of production data copied to test environment
- Anonymized production: Tool(s) used to perform irreversible encryption, substitution, or scrambling of values
- Pseudo-anonymized production: Reversible anonymization
- In this presentation, I'll address factors to consider when choosing these options
- At the enterprise level, you'll probably need one or more tools
- More than one option might be useful when creating test data
- Whatever options are chosen, be really careful about:
  - Testing on actual production data in the production environment
  - Copying test data into the production environment

Test Data Is Simple When Life Is Like…

- So, if you are testing a simple mobile app, you might be thinking, "Test data, what's the big deal?"
- If the production data is small enough, you can generate data to populate the database
- Or maybe there's no personal information in the production data, so you just use that
- In some cases, your testing situation might simple, like the one shown here, but eventually it could evolve as your application's abilities grow



But the Enterprise Be Like…



Actually, it's usually more complicated than this…

Production Data, the Testing Gold Standard

- For most large-scale systems, testers cannot create sufficient and sufficiently diverse test data by hand
- Data generation tools can create large datasets, but the data might lack diversity, similar value distributions, or other attributes of production data
- Production data is ideal, especially when large data sets have accumulated over years with various current and past application versions
- Production data looks exactly like the data the application will encounter on release, since it is
- But using production data for testing poses many challenges

Privacy Concerns: Fly Meets Ointment

- Production data often contains personal data
- Trying to secure test data handling during testing can impose undesirable inefficiencies and constraints
- Such inefficiencies and constraints apply especially when using distributed and outsourced testing
- Enter the anonymization tool
- However, such tools are not magic
- Misuse can result in trivially easy reversal
- For example, letter substitution could lead to names like "Kpio Cspxo," obviously "John Brown"
- In addition, Kpio Cspxo (and other nonsense) is not realistic
- For testing you must preserve data meaning and meaningfulness

Meaning and Meaningfulness

- Substitutions, good and bad
  - Good: "John Brown" to "Lester Camden"
  - Bad: "John Brown" to "Charlotte Dostoyevsky" (esp. if we have gender field)
- Queries, views, and joins of anonymized data must yield same results as on production data
- For example, if a production query on "John Brown" returned 20 records, a test data query for "Lester Camden" must also return 20 records
- Problems here will affect functionality, performance, reliability, and load testing
- In addition, localization can be an issue with different languages and cultures, especially if multiple character sets can apply

Oh, Those Long-Distance Relationships

- Data and relationships between data can span databases
- For example:
  - Three applications have years of data describing the same population, spread across three different databases
  - The applications interoperate and share data
  - Data-warehousing and analytical applications access the related data across databases
  - Logical records are created with de facto joins across via de facto foreign keys
  - Anonymization can break these integrity constraints
  - Meaningful end-to-end testing of functionality, performance, throughput, reliability, localization, and security becomes impossible
- For many of our clients, the usefulness of the test data for interoperability testing poses the hardest challenge
- This is true whether hand generated, tool generated, production, pseudo-anonymized, or anonymized

Generic Test Data Requirements

- Keep existing data quality levels
  - Most production data contains a large number (~25%) of errors
  - Error rates must be the same for test data, but, if privacy is an issue, must not enable reverse engineering of the data
- Ensure maintainability
  - Be able to edit, add, update, and delete data
  - This includes individual fields and records, whole database, and across logical records spanning databases
- Maximize efficiency of the test data acquisition and update processes
- If production data is to be copied and/or anonymized, ensure those operations occur on quiescent data

Test Database Server Costs

- Yet another challenge to production-scale test data is cost
- The test data must reside on test database servers, which can involve:
  - DBMS licensing costs
  - Server costs
  - Disk space costs
  - Ongoing hardware and software maintenance
  - A regular backup process (with attendant costs for backup media as well as labor)
- Using open-source DBMS only addresses the first of these costs

Storage Virtualization

- This allows abstraction of physical storage space
- It can reduce some of the cost issues associated with test data storage
- It's probably under-utilized by testing groups, as I hear very little about it online, from clients, or at conferences
- It can help in terms of allocating/de-allocating production-like storage for testing
- The issue of anonymization of the production data still exists, though

Coping with Data Size

- Enterprise data have gotten huge, due to increases in disk capacity
- Unfortunately, disk throughput has not kept up
- This creates serious issues with copying, anonymizing, and managing test data that comes near to or achieves production size

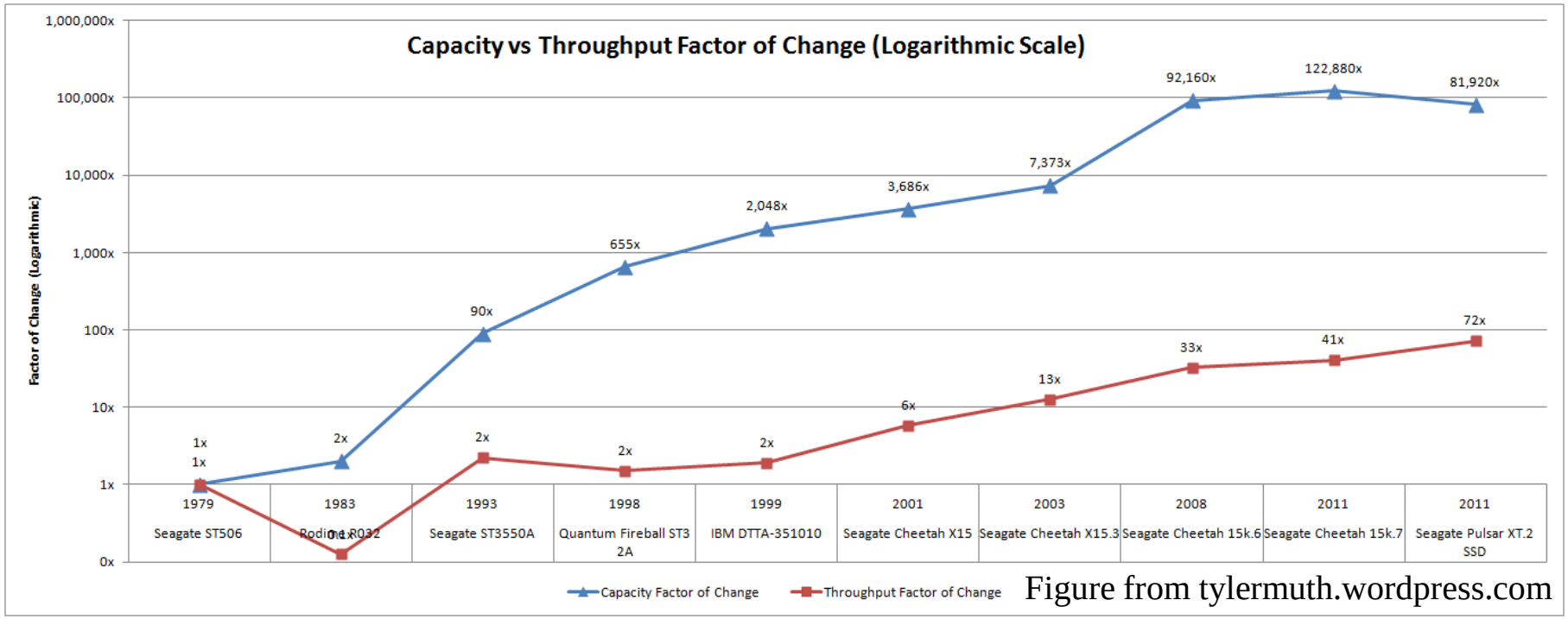


Figure from tylermuth.wordpress.com

Specialized Application Support

- Many enterprises use tools such as SAP to help manage their businesses
- There are test data anonymization tools for the more common such enterprise applications
- If you are using a more obscure application, no such tool may be available
- Similar issues exist with NoSQL and other data formats
- Further, the application vendor might not be willing to share metadata information
- This should be a consideration during enterprise application selection
- However, testers are rarely included in such decisions, and once made, such decisions often don't change for years

Success with Test Data Tools

- Success with test data tools is like any other test tool
  - Assemble a team
  - Elicit requirements from users and stakeholders (including those who understand privacy/security issues)
  - Identify risks and constraints
  - Determine tool options (open-source and commercial)
  - Evaluate tools to identify short-list options
  - Select the tool
  - Pilot the tool
  - Deploy the tool
- Short-cutting this process or imposing unrealistic budget targets can cause significant inefficiencies and possibly even inability to fulfill basic needs

Conclusions

- Testing enterprise applications poses challenges beyond what small-scale applications pose
- This is especially true for test data
- Challenges include sources of data, privacy, availability of tools, size, cost, and testing usefulness
- Like any complex situation, success requires understanding the challenges and planning to meet them
- This presentation should help you identify your specific challenges and build an appropriate plan

…Contact RBCS

For over twenty years, RBCS has delivered software and hardware testing services, including consulting, outsourcing, and training. Employing the industry's most experienced and recognized consultants, RBCS conducts product testing, builds and improves testing groups, and hires testing staff for hundreds of clients worldwide. Ranging from Fortune 20 companies to start-ups, RBCS clients save time and money through improved product development, decreased tech support calls, improved corporate reputation and more. To learn more about RBCS, visit www.rbcs-us.com.

Address:	RBCS, Inc. 31520 Beck Road Bulverde, TX 78163-3911 USA
Phone:	+1 (830) 438-4830
E-mail:	info@rbcs-us.com
Web:	www.rbcs-us.com
Facebook:	RBCS, Inc
Twitter:	@RBCS, @LaikaTestDog