Realistic Synthetic Data @ Scale
Influenced By, But Not Production Data

Mehul Sheth
Performance Engineer
MEHUL SHETH
Performance Engineer

GLASS IS HALF FULL / EMPTY??

TO A PERF. ENGINEER
GLASS IS 50% UTILIZED !!!
**PROD ISSUE**
Alert
▲ Memory
⊗ Response
Crash

**BLAME GAME**
Our Env. ✓
QA ✓
Perf. ✓

**ROOT CAUSE**
Unseen data
Edge Case
⊗ Prod Data
CRITERIA: Define data, impact product behavior

MODEL: Analyze prod. data, model characteristics of production

APPLY: Create realistic test data at scale
1. BETTER INSIGHTS
   Variation in production data characteristics

2. BETTER DECISIONS
   Architecture decisions based on facts rather than assumption

future
details

about
problem
solution
benefits
DEVIL IS IN THE DETAILS
DATA SET

A folder containing files and folders

CRITERIA

Folder hierarchy
Size and type of files
Compressibility of files
Density in folders

CHALLENGES

Accessing correct data
Valid and consistent data
Generating data larger than available
1. Production Data Modeling

- Walkthrough production data
- Extract relevant information
- Identify, Parse, Transform & Aggregate

2. Synthetic Data Generation

- Use the generated model
- Define size of data, Create arrays (iterative)
- Generate data as per arrays
Production Data Modeling
Identify relevant data
Production Data Modeling

Representation of model

DIRS_PER_DIR = \{0.5:(0,1), \ldots\}
FILES_PER_DIR = \{0.2:(0,3), 0.5:(3,5), 0.75:(5,12) \ldots 1:(60,401)\}

Parameter

Cumulative probability
Synthetic Data Generation
Text Data Generation Algorithms

Random Text
Simple, Fast, Garbled, only size is useful
k6K8ETXiREXqh3pX, llp0nlsqM43

Obfuscated using Caesar’s Cipher
Simple, Fast, replace char by another, need corpus
man>3=pdq, zebra>5=ejgwf, cipher>20=wcjbyl
Synthetic Data Generation
Text Data Generation Algorithms

Statistical model-based generation
Large Corpus, Observation distribution saved with probabilities
THE, THIN, THIS

Markov model-based generation
Parse public data (like Wikipedia), create probabilities using unigram, bigram or trigrams.
Random Bytes
- Pack random number into bytes
- Range of random number decides compressibility
- Stack all bytes to get the binary file
Other data
Mails, DB, etc.

Bigger scale
Petabytes of data

Signature
Files with matching signature and contents
Thank You!

Please take a moment to rate this session.

Your feedback matters to us.