The Power of Random Identifiers

Derk Norton
Engineering Manager
Crater Dog Technologies™

November 16, 2013

What are Random Identifiers?

- They are like *normal* identifiers
  - SS#: 123-45-6789
  - Driver's License: CA-12-345-6789
  - Credit Card: 0123-4567-8901-2345
  - Email: derk.norton@gmail.com
- But *much* better!
  - contain no sensitive information
  - can't be guessed
  - no risk of collisions with other identifiers

What Do They Look Like?

- They are really just character strings
  - base 32 encoded
  - 0..9, A..D, F..H, J..N, P..T, V..Z (missing E,I,O,U)
  - example: `QM65JYZ0JD2245A43C8YWQV4D2V1Y8L6`
- But underneath lies the magic
  - derived from very large cryptographically random numbers
  - a very nearly random distribution across all possible values
  - for example, 16 byte numbers give 2^128 possible values!

Just How Many Values is That?

- It's almost unimaginable
  - 340,282,366,920,938,463,463,374,607,431,768,211,456
  - 2^128 ≈ 3.4E38
- More than the number of grains of sand in our galaxy
  - the number of grains of sand on earth ≈ 1.0E25
  - the number of stars in our galaxy ≈ 1.0E12
  - if we assume 1 sandy planet per star
  - 1.0E25 x 1.0E12 = 1.0E37 grains of sand in our galaxy
- Roughly the number of atoms in all humans on Earth
  - atoms in your body ≈ 1.0E28
  - the number of humans on Earth ≈ 1.0E10
  - 1.0E28 x 1.0E10 = 1.0E38 atoms in the human race

Chances Are...

- Extremely unlikely the *same* value will occur twice!
- Why? $$P(duplicate) = 1 - \frac{N!}{N^n (N-n)!} \approx \boxed{1 - e^{-n^2/2N}}$$
  - let's say there a billion billion things you want to identify
  - n = 1.0E18 is your number of samples
  - N = 3.4E38 is the total number of possible values
  - P(duplicate) ≈ 0.0015 across all samples
- Anyone want to place a bet on a collision?

Choosing an Identifier Size

Scope	Number of Things	Number of Bytes	Number of Bits	Number of Characters	Total Size of Space
Desk	16	1	8	2	256
Room	256	2	16	4	65536
Building	65536	4	32	7	4.29E+009
City	4.29E+009	8	64	13	1.84E+019
Globe	1.84E+019	16	128	26	3.40E+038
Galaxy	3.40E+038	32	256	52	1.16E+077
Universe	1.16E+077	64	512	103	1.34E+154

- The *total space size* should be at least the **square** of the **maximum** *number of things* to be uniquely identified.
- This results in a probability of an identifier collision across **all** things of less than 39%.

What Does This All Mean?

- Unguessable Identifiers
  - if I can't generate a duplicate, neither can anyone else
  - no need to encrypt since they are already random
  - less chance of identify theft
- No Database Collisions... Ever
  - no need for centralized key generation
  - no need to check for duplicates
  - keys are self-hashing (very efficient!)
  - allows easy splitting and merging of databases