

A Storage Engine for Amazon S3

MySQL Conference & Expo 2007

Mark Atwood <mark@fallenpegasus.com>

<http://fallenpegasus.com/code/mysql-awss3>

Storage Engines are Protocol Translators

- They transform some other data protocol and some other data model into MySQL's presentation of a db, tables, rows, etc
- Traditional Storage Engines translate filesystem and block i/o system calls
- Network Storage Engines
 - Instead of copying the network into your database, make the network look like a database!
 - Examples: Federated, ODBC, HTTP, MemCacheD, AWS S3

What is S3?

- Amazon Web Services Simple Storage Service
- HTTP/REST based protocol
- Simple data model: buckets, items, item keys, item contents, all with ACLs

“Storage for the Internet.”

Buckets

- Global namespace
- If a bucket's name looks like a domain name you can access it at that domain name.
- Contains "Items"
 - a few items
 - a billion items, or more

Items

- Key: looks kind of like a URL path, if you want it to
- Type: MIME Type
- Small amount of user metadata: name1=value, name2=value, etc
- Contents: 1 byte, to 5 GiB, of MIME type'd data
- billions of items per bucket is no problem
- 100s of TiB is no problem

Hundreds of terabytes

Virtual Hosting of Buckets

- name a bucket “s3.example.com”
- CNAME “s3.example.com” to “s3.amazonaws.com”
- set ACL of bucket & items to “public-read”
- uses Amazon's internet bandwidth and datacenters
- cheaper than edge distribution, and works on all your static or slowly changing data

Cost

- 15¢ month GiB storage
- 20¢ GiB xfer
- xfer to EC2 is free
- competition is on the way

Owning your own disks is painful

- because you have to buy it before you use it
- buy with capital, not as an expense on revenue
- "An empty disk costs as much as a full one."

An empty disk costs as much as a full one.

- You **still** don't really own a durable capital good
 - your disks are going to be painfully small in 3 years, and scrap in 6
 - your raid enclosures are going to be EOLed
 - you are going to have to spend all that money **again**
- Datacenters are painful too
 - they cost capital, and a lot of it
 - cost of renting rack space, by the cubic inch
 - cost of power and cooling. There is a reason that they're getting built next to dams!
 - cost of monitoring and maintenance
- RAID only makes hw failure survivable, it doesn't make it affordable
If you can't afford to keep your raid fed, you can't play the game.

Disaster Recovery & Geographic Distribution

- what do you do when your datacenter burns down, falls down, blows up, washes into the ocean?
- Buy another data center?!
 - you've spend all the money **again**
 - plus the pain of geographic data replication and synchronization

Translating the S3 data model to the RDBMS data model

- An AWS account becomes a CREATE SERVER command
- An S3 bucket becomes a table
- An S3 item becomes a row
- An S3 item key becomes a primary key
- An S3 item contents becomes a BLOB or a VARCHAR

Example SQL statements

```
CREATE SERVER 'MyAWSacct'  
  FOREIGN DATA WRAPPER 'AWS'  
  OPTIONS  
    (USER 'aws id string',  
     PASSWORD 'aws secret string');  
  
CREATE TABLE 'bierce'  
  ('word' VARCHAR(255) NOT NULL PRIMARY KEY,  
   'defn' BLOB)  
  CHARSET=utf-8  
  ENGINE=AWSS3  
  CONNECTION='aws3 DevilDictionary $server MyAWSacct';  
  
SELECT defn FROM bierce WHERE word='WIT';  
  
INSERT INTO bierce (word, defn) VALUES  
  ('AUTHOR',  
   'One noted for confusing bitterness with humor.');
```

```
DELETE FROM bierce WHERE word='AUTHOR';
```

Use cases

- Saving EC2 work
- SQL CMS via S3 virtual hosting
- Huge list of persistent primary keys
- Big slow blobs, to join against fast local tables
- "The Image Server Problem"
- Warm storage of large datasets

Not Transactional

No Temporal Guarantees

- Write replication can be delayed and reordered
- Cluster folks might recognize this

The Future

- Code Improvement
- S3 & HTTP metadata
- Multiple Data Columns
 - Sharing a solution with the HTTP and the MemCacheD engines
- Information Schema
- Security and Authentication
 - TLS, X.509 certs
 - Alternative credential storage
- Transfer & Storage Compression
- More Storage Engines
 - A storage engine for AWS SQS
 - More AWS services & AWS competitors
 - Cluster storage & Replication distribution via S3
 - The holy grail for AWS EC2 users:

"A **Generic Schema** Storage Engine for Amazon S3"

Availability

- MySQL 5.1
- GPL
- <http://fallenpegasus.com/code>