METADATA AT DROPBOX: A LOOK AT DROPBOX’S TRANSACTIONAL DATABASES

PETER BOROS
CS3 AMSTERDAM, 2017
ABOUT ME

MYSQL SRE @ DROPBOX
OUR GLOBALLY DISTRIBUTED TEAM

DUBLIN, SAN FRANCISCO, SEATTLE & MALTA
DROPBOX VISION
FOUNDED IN 2007
FILE, SYNC & SHARE
What we’ve built

500 million people around the world use Dropbox to work the way they want, on any device, wherever they go. With 200,000 businesses on Dropbox Business, we’re transforming everyday workflows and entire industries.

3,300,000,000
sharing connections have been created with Dropbox

1,200,000,000
files are saved on Dropbox every day
NEXT FEW YEARS
2016 HIGHLIGHTS

MAGIC POCKET, INFINITE, PAPER
EUROPA

STORING DATA AND METADATA IN EUROPE
DATABASE INFRASTRUCTURE

LAYING THE FOUNDATIONS
IMPROVING THE FLYING ROCKET SHIP
SETTING OUR DATABASE TEAM UP FOR SUCCESS
WHERE ARE WE HEADING NEXT?
LAYING THE FOUNDATION

WHERE DID IT ALL BEGIN?
WHERE WE STARTED

Example #1: High-level architecture

- early 2008
- 4 employees
- (2 backend)
- 50k users

KMOD ON SCALE
LET’S USE MYSQL

METADATA STORAGE
OUR FIRST DATABASES

SERVER FILE JOURNAL
GLOBAL
SERVER BLOCK CACHE

STARTING LEAN
IMPROVING THE FLYING ROCKET SHIP
BIGGEST DATABASES TODAY
SERVER FILE JOURNAL
EDGESTORE
MAGIC POCKET
AND 10S OF SMALLER ONES
We built our own system that would abstract away the database by providing higher-level abstractions and use MySQL (InnoDB) as the storage engine.
HIGH PERFORMANCE AND SCALABLE STORAGE

Edgestore: 2012 — Present
EDGESTORE ENTITIES & ASSOCS

GRAPH-LIKE DATA MODEL
THE MIGRATION TO EDGESTORE

Metadata
MySQL  MySQL  ...  MySQL

Several Petabytes
1000s of MySQL shards

SERENGETI
MAGIC POCKET

IMMUTABLE BLOCK STORAGE
TRY TO KEEP MACHINES HOMOGENEOUS

OUR MACHINE CLASSES ARE NAMED AFTER WWE WRESTLERS
SCALING OUR DATABASES

1 million users
7 months after launch

1 million files
every 15 minutes

500 petabytes
stored in Magic Pocket

YEARS

‘07 ‘08 ‘09 ‘10 ‘11 ‘12 ‘13 ‘14 ‘15 ‘16
DAYS SINCE LAST MACHINE FAILURE

0

MASTER PROMOTION HAS TO BE VERY RELIABLE
DBMANAGER: JOB EXECUTION

Running 1000s of jobs every day.

<table>
<thead>
<tr>
<th>Job</th>
<th>Status</th>
<th>Timeline</th>
<th>Attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td># Clone 30229-3248 to 30229-3249 to 30229</td>
<td>Completed</td>
<td>2016-09-12 14:50:03</td>
<td>0</td>
</tr>
<tr>
<td># Clone 30229-3248 to 30229-3249 to 30229</td>
<td>Completed</td>
<td>2016-09-12 14:50:03</td>
<td>0</td>
</tr>
<tr>
<td># Clone 30229-3248 to 30229-3249 to 30229</td>
<td>Completed</td>
<td>2016-09-12 14:50:03</td>
<td>0</td>
</tr>
<tr>
<td># Normal Prometheus slow</td>
<td>Completed</td>
<td>2016-09-12 14:50:03</td>
<td>0</td>
</tr>
<tr>
<td># Normal Prometheus slow</td>
<td>Completed</td>
<td>2016-09-12 14:50:03</td>
<td>0</td>
</tr>
</tbody>
</table>
A DROPBOX STORY: SWEATING THE DETAILS

IMPORTANT EARLY DECISIONS

SSDs
A DROPBOX STORY: SWEATING THE DETAILS

INNODB COMPRESSION

PERF BENCHMARKS, RELIABILITY TESTS & DATA COMPARISON TESTS
A DROPBOX STORY: SWEATING THE DETAILS

GTID ROLLOUT

CANARY TO PROD
GTID

mysql-bin.002345 161

mysql-bin.008734 1274  mysql-bin.000845 234535

UNIVERSAL POSITIONING ACROSS ALL DATABASE HOSTS
UNIVERSAL POSITIONING ACROSS ALL DATABASE HOSTS
A DROPBOX STORY: SWEATING THE DETAILS

INNODB
PAGE SIZE

PERF BENCHMARKS, RELIABILITY TESTS & DATA COMPARISON TESTS
SWEATING THE DETAILS: 4K PAGES

USEFUL IF TYPICALLY SMALL PORTION OF PAGES ARE ACCESSED
DURABILITY

SEMI-SYNC REPLICATION WITH VERY LARGE TIMEOUT

WE DON’T TRUST OR COUNT ON MACHINE LOCAL DURABILITY
STORAGE

RAID0

FAIL FAST
LEARNING FROM FAILURE

DISTRIBUTED STATE VERIFICATION
& FASTER DISASTER RECOVERY
WE WORK HARD TO UNCOVER SYSTEMIC WEAKNESSES

ARE YOU SURE EVERYTHING WILL BE OK?
HUMAN ERRORS?

```
gsh -q hwclass=valentine lifecycle=repair "sudo upgrade_db"
```

MORE NOT ANTIFRAGILE TOOLING
HUMAN ERRORS?

gsh -q "hwclass=valentine lifecycle=repair" "sudo upgrade_db"

THE CORRECT COMMAND
gsh -q "hwclass=valentine lifecycle=repair" -- "sudo upgrade_db"

HUMAN ERRORS?

EVEN MORE CORRECT COMMAND
ANTIFRAGILE AUTOMATION

HAPPY PANDA AT WORK
DROPBOX CASE STUDY:
SETTING OUR DATABASES
TEAM UP FOR SUCCESS

#GOTEAM
WE NOT I

WE VALUE THE STRENGTHS OF EACH PERSON ON OUR TEAM
TEAM TRADITIONS

WE MAKE THE MOST OF OUR TIME TOGETHER
BUILD CONFIDENCE

WE TEST OUR SYSTEMS AND TOOLS
DATABASE CONTAINER

EASY INTEGRATION TESTING
OUR WORK IN 2016 GREATLY REDUCED TIME SPENT ON SEVS

WE NOW HAVE MORE TIME TO BUILD TOOLS
DROPBOX CASE STUDY: WHERE ARE WE HEADING NEXT?

2017 AND BEYOND
MYSQL 5.7

MANY NEW USEFUL FEATURES FOR DROPBOX
REPLICATION ENHANCEMENTS

slave_parallel_type = LOGICAL_CLOCK

(old behaviour DATABASE)

rpl_semi_sync_master_wait_point = AFTER_SYNC

(old behaviour AFTER_COMMIT)
GROUP REPLICATION

SYNCHRONOUS
AUTOMATION & TOOLING

A human operator must authorize execution here, unless the Prescription was instantiated with run_automatically=True.

IssuePullPlugin --- > DiagnosePlugin --- > RemediatePlugin --- > IssuePullPlugin

Hooks

Issues

Issue

Prescription

Hooks have the ability to postpone execution or even to completely deny it.

INCREASING THE INTELLIGENCE OF AUTOMATION
SHAMELESS PLUG

Coffee break

Amsterdam, SURFSara  
10:40 - 11:40

Automated error handling at Dropbox  
Karoly Nagy et al.

Amsterdam, SURFSara  
11:40 - 12:20

BY MAXIM BUBLIS AND KAROLY NAGY
SHARING OUR JOURNEY

Improving the performance of full-text search

Adam Bechhofer  ·  September 7, 2016

Initially, Dropbox's full-text search engine was always less than optimal. The slow response times were an embarrassment for us. I'd like to show you the steps we took to improve our search latency from 250 ms to 120 ms (95th percentile). We knew what we wanted to do—we measured, we analyzed, we fixed.

Problem

In order to create a good user experience, we strive to keep our query latency under 250 ms (95th percentile). We noticed that our latency had deteriorated quite a bit since we started adding users to the system.

Read more →