Managing Your Databases and Sharding with Jetpants

Bob Patterson Jr Database Engineer @ Tumblr MySQL Connect 2013



Background

What is Jetpants?



What is Jetpants?

- → MySQL management suite
 - → Command Line Tool
 - → Ruby automation library
- Developed by Tumblr (created by Evan Elias)
- Open sourced in June 2012 https://github.com/tumblr/jetpants

MySQL at Tumblr



MySQL at Tumblr

- → 250+ machines
- → 54 Terabytes
- → 229,208,178,189 rows of data
- → 3 engineers

Dividing the Data

What is Sharding?

Sharding is...

- → Logically Dividing your data
- → Horizontally Scaling

Why Shard?

Reasons to Shard

- → Scale writes
- → Eliminate single points of failure
- → To much data for one machine

Sharding Options

- → Functional Partitions
- → Dynamic Range-Based
- → Static Range-Based
- → Hash / Modulo
- → Lookup Table

Functional Partitions

- Each pool contains
 - → 1 Master
 - → 0-x Active Slaves
 - → 2 Standby slaves
- → Global Tables / Unshardable Tables

Range Based Sharding

Ranged Based Sharding

- min and max id range
- → Ranges can be any size
- → Easy to generate config file
- → Each Shard Contains
 - → Same tables
 - → Different sliver of the overall dataset

Ranged Sharding at Tumblr

Data moves around



Data moves around

- → Shards can be split into N new shards
- → Splits are non-disruptive
- → Shards can me merged to join under-utilized shards (new feature)

Benefits of Ranged Sharding

Benefits of Ranged Sharding

- → Low shard count
 - we do not pre-allocate thousands of small shards
- Shards created on demand
- → Shard splits (and now shard merges) defrag the data

Pitfalls of Ranged Sharding

Pitfalls of Ranged Sharding

- → Uneven load on your shards
- → Uneven usage of disk capacity



Jetpants Installation

Jetpants Installation

- → Requires ruby 1.9.2+
- gem install jetpants
- May need to install system packages first (mysql-devel)

MySQL Setup

- → MySQL / Percona Server 5.1 or 5.5 (5.6 soon)
- → InnoDB / Percona XtraDB (need clustered index)
- → First column of PK on sharded tables is sharding key
- → No auto_increment tables on shards
- One mysqld and one database per server

Jetpants Command Line Tools

Jetpants Command Line Tools

- jetpants pools
- jetpants summary
- jetpants clone_slave
- jetpants shard_split
- jetpants promotion
- jetpants shard_merge (coming soon)

JETPANTS COMMAND LINE

jetpants clone slave

jetpants clone slave

- → Utilizes a standby slave
- → Stops mysqld and copies over files
- → Uses tar, nc and compression with apress (you can set your own compression)
- Can create multiple slaves by chaining them with tee and fifo on intermediate nodes (uses full up and down link)

jetpants exporting data

jetpants exporting data

- Divide data by primary key into n chunks
- → Spawn multiple threads to export data
- → Each thread does a SELECT INTO OUTFILE

jetpants importing data

jetpants importing data

- → Spawn multiple threads again
- → Threads do LOAD DATA INFILE
- Disables binary logging before the import
- → Also sets innodb-autoinc-lock-mode = 2
- → Disable unique_checks

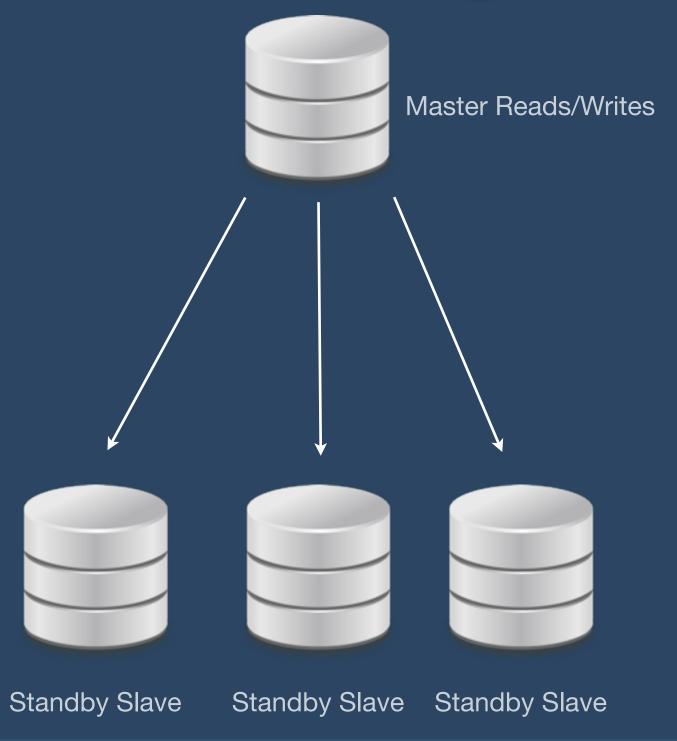
jetpants removing masters

jetpants removing masters

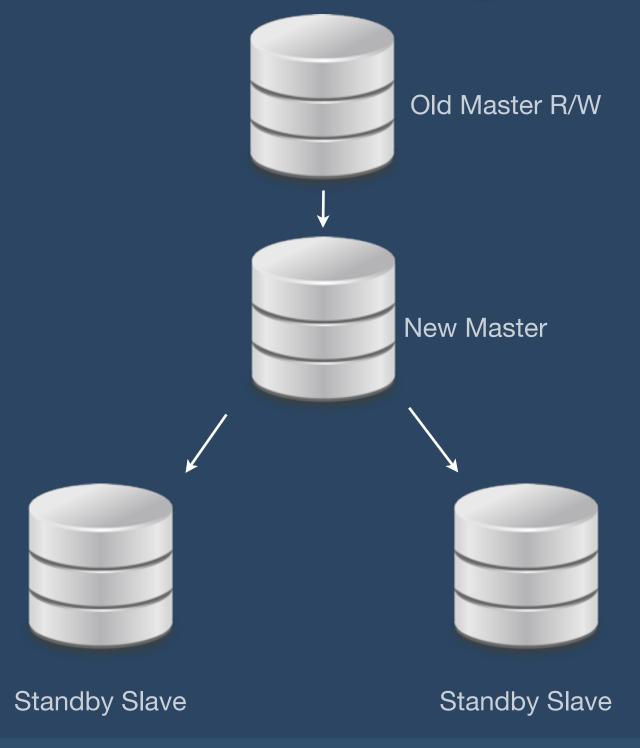
- → Lockless
- → Point reads to new master
- Point writes to new master
- → No auto_increment tables

JETPANTS COMMAND LINE

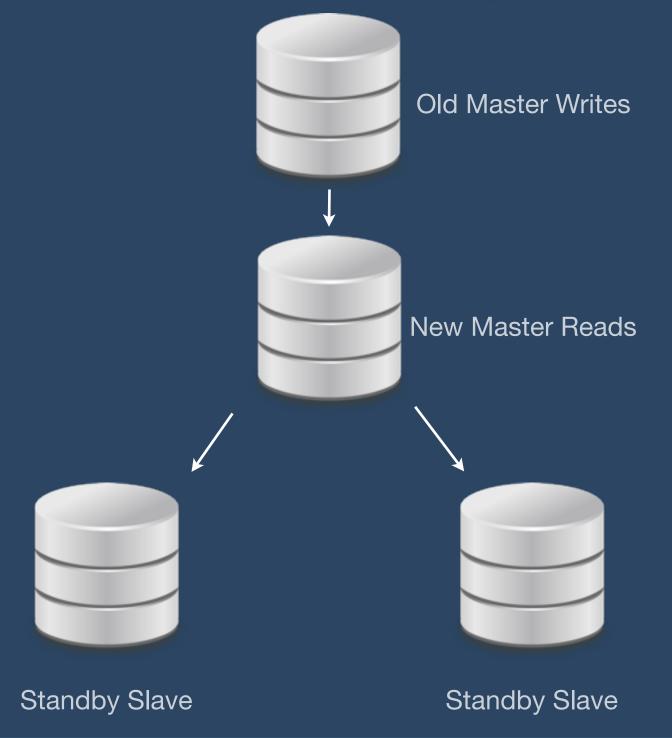
jetpants removing masters



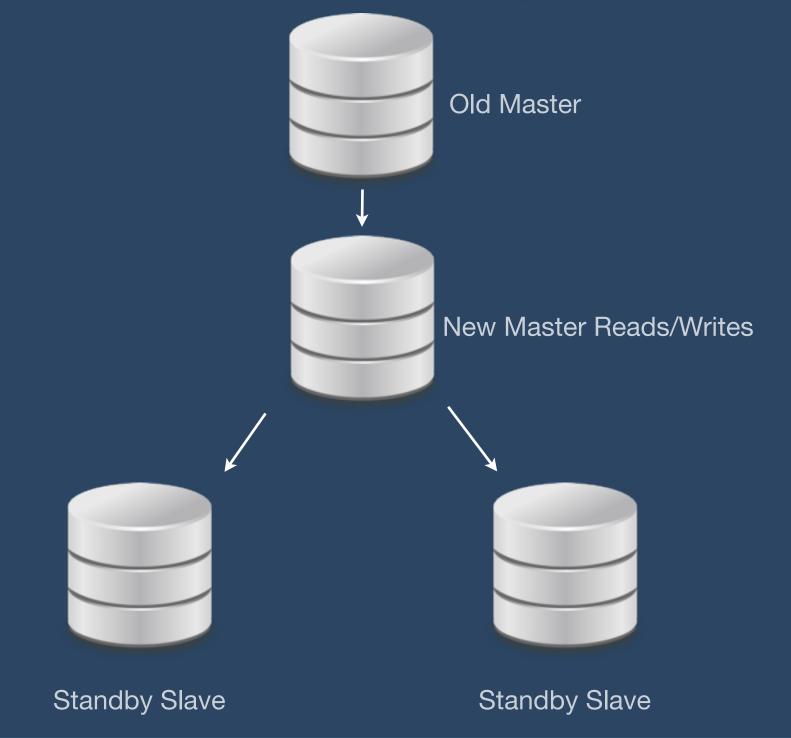
jetpants removing masters



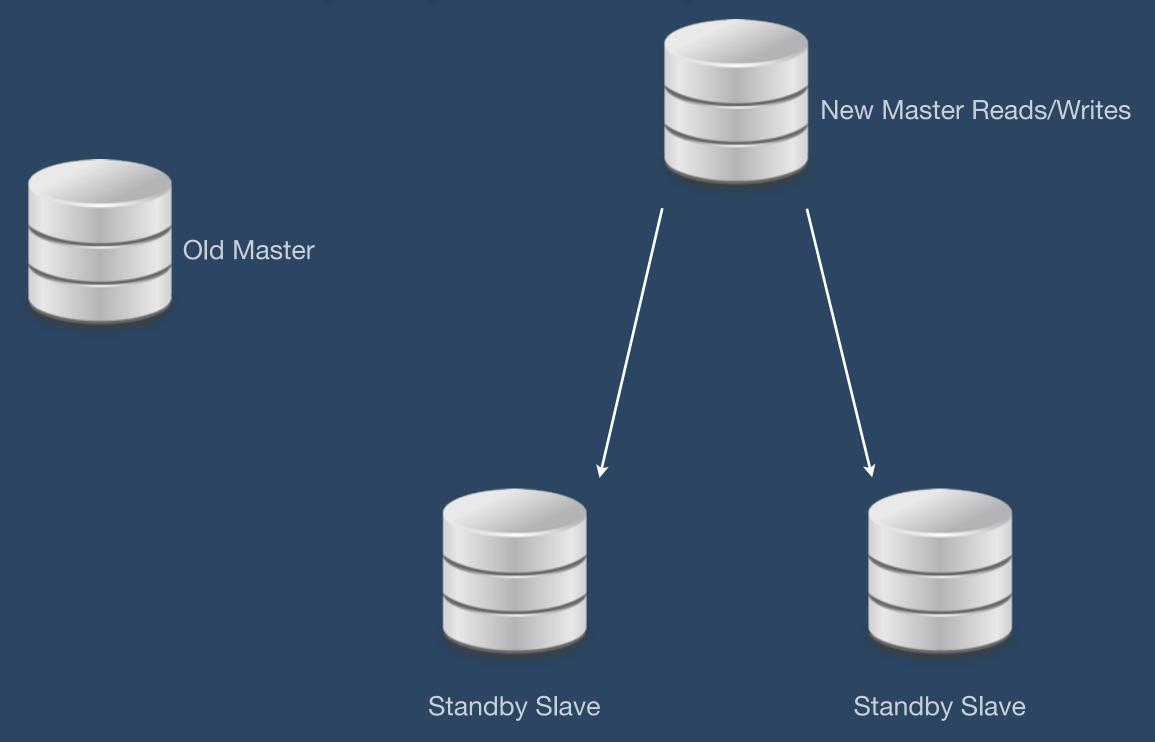
jetpants removing masters



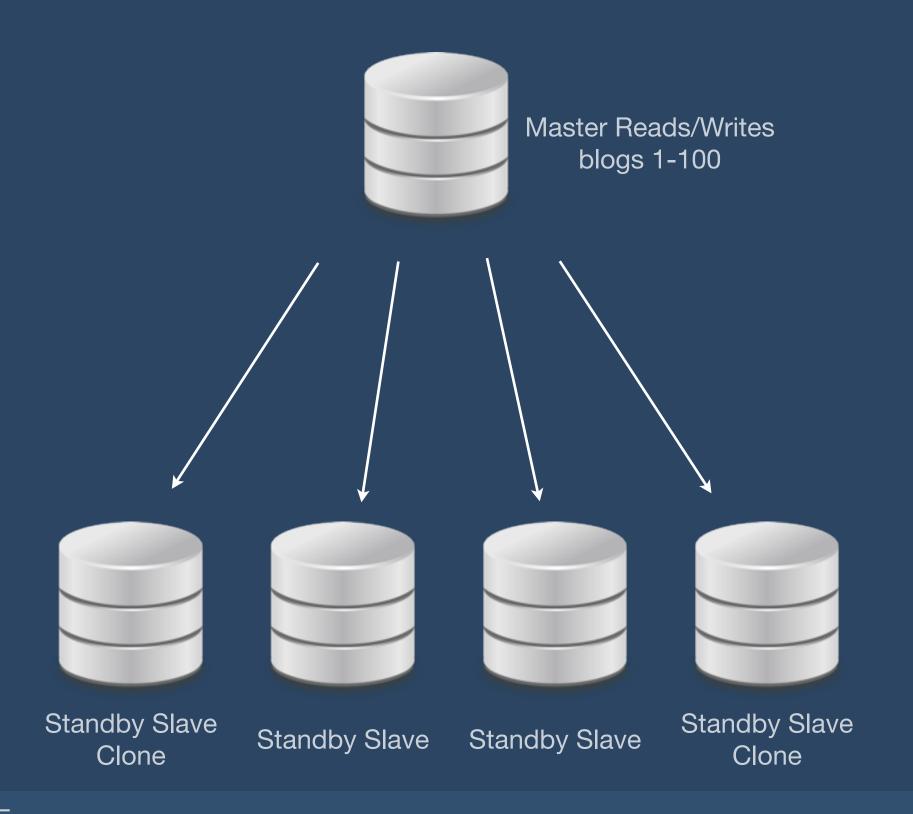
jetpants removing masters

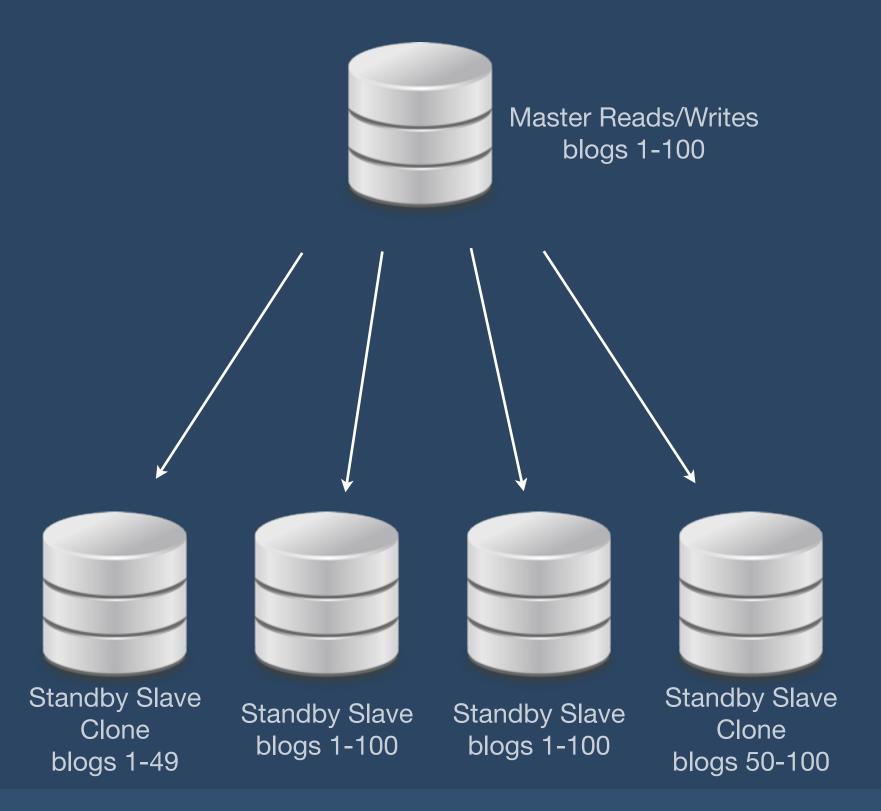


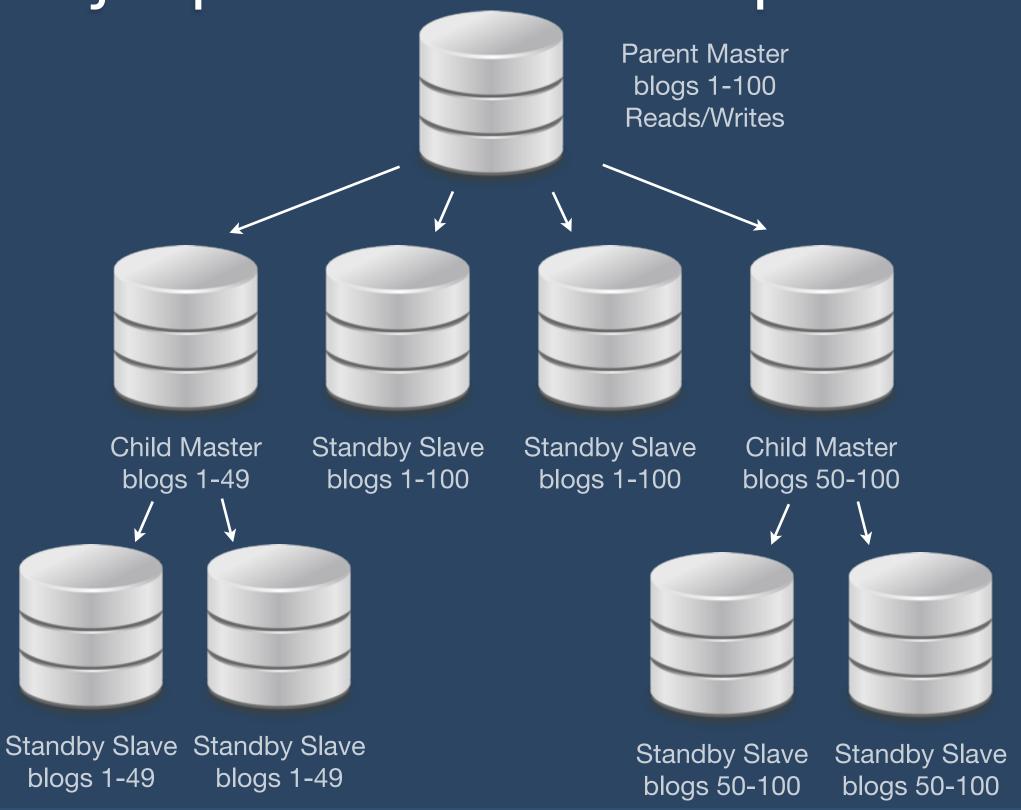
jetpants promotions

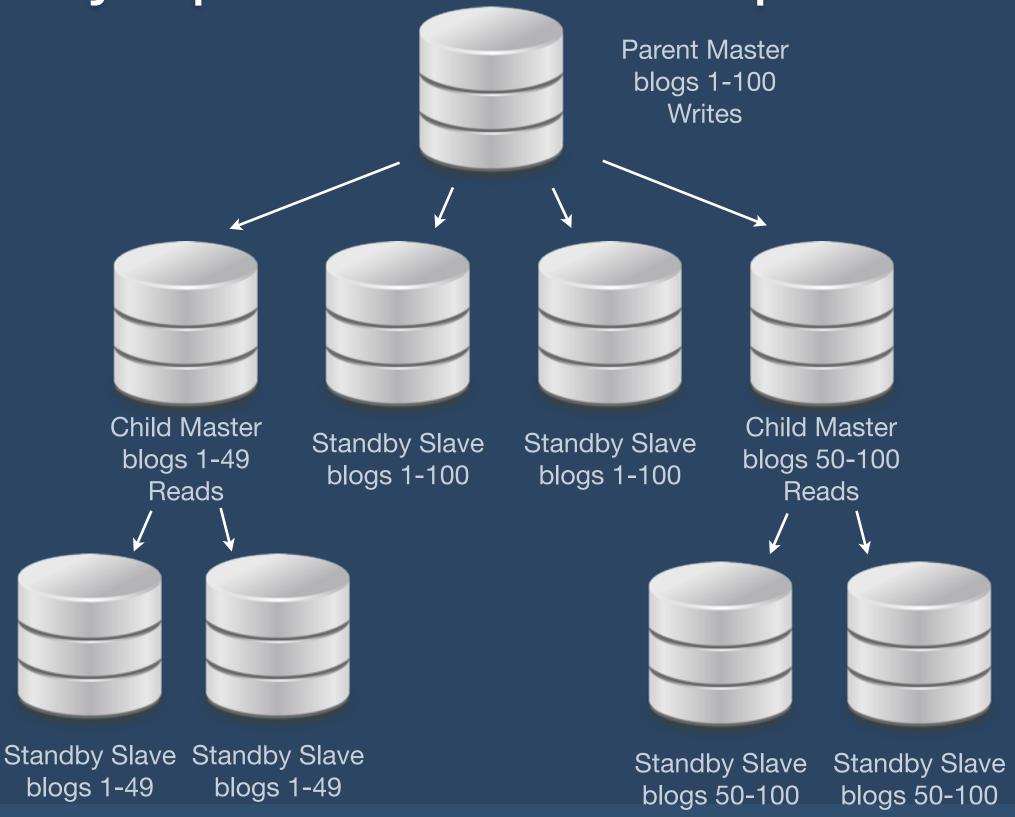


- → Clone one new slave for each new shard
- → Export dataset that will live on new shards
- Drop the tables and recreate tables
- → Import data
- → Replay replication stream from parent shard
- → Remove data from other shards













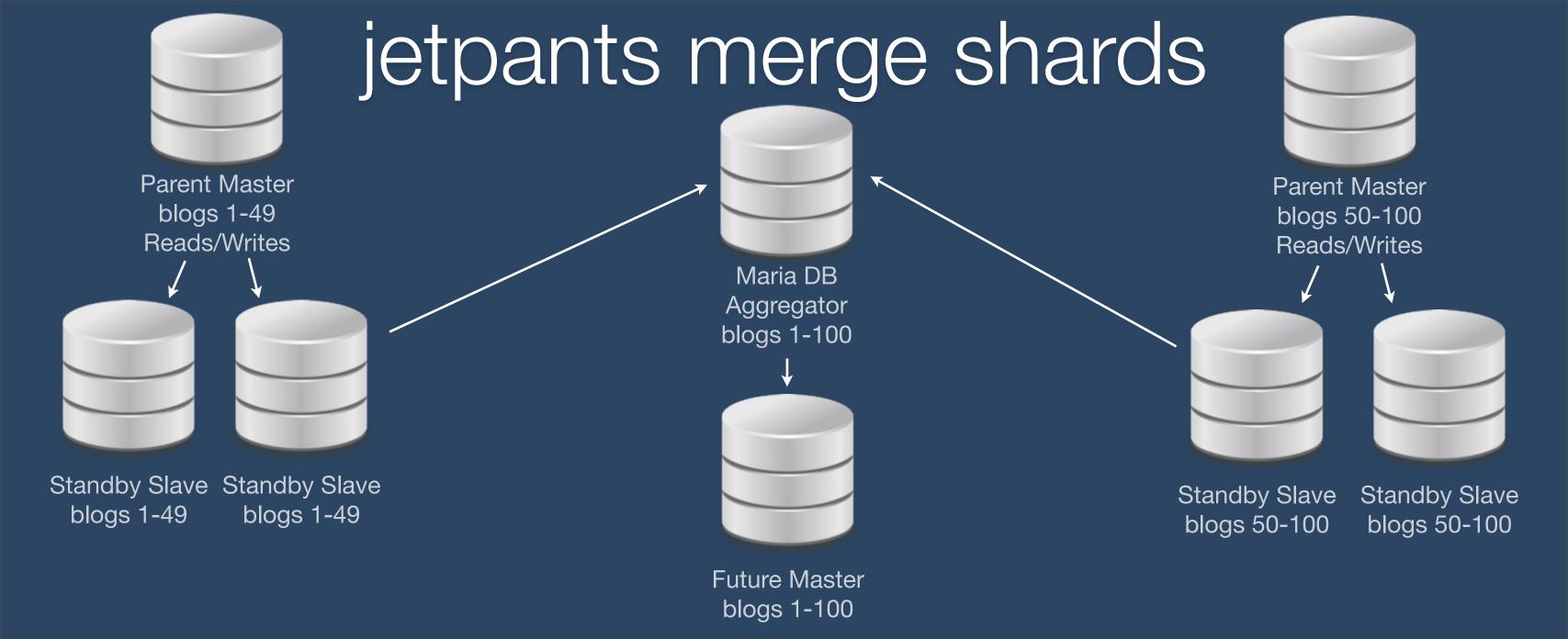


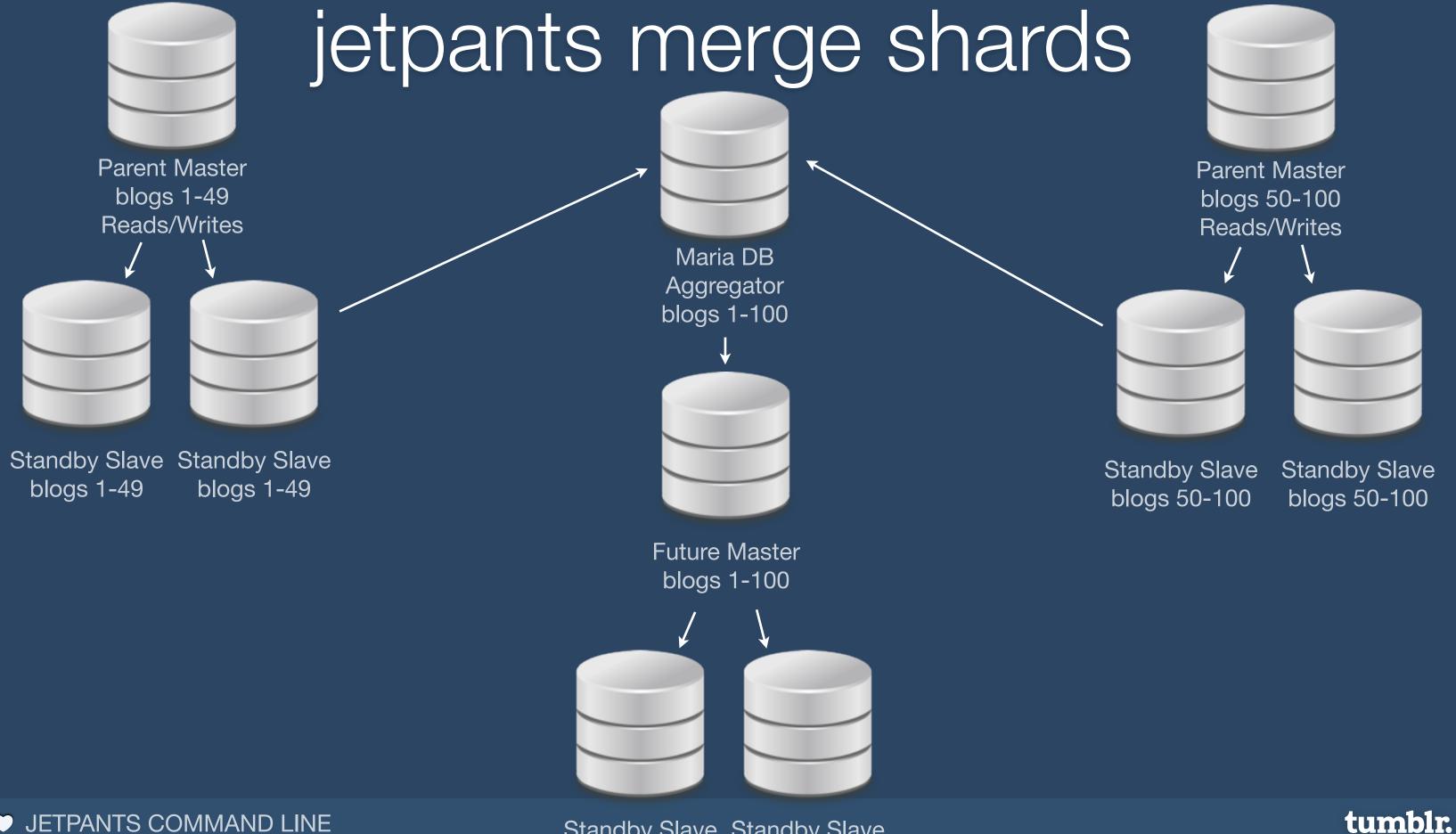


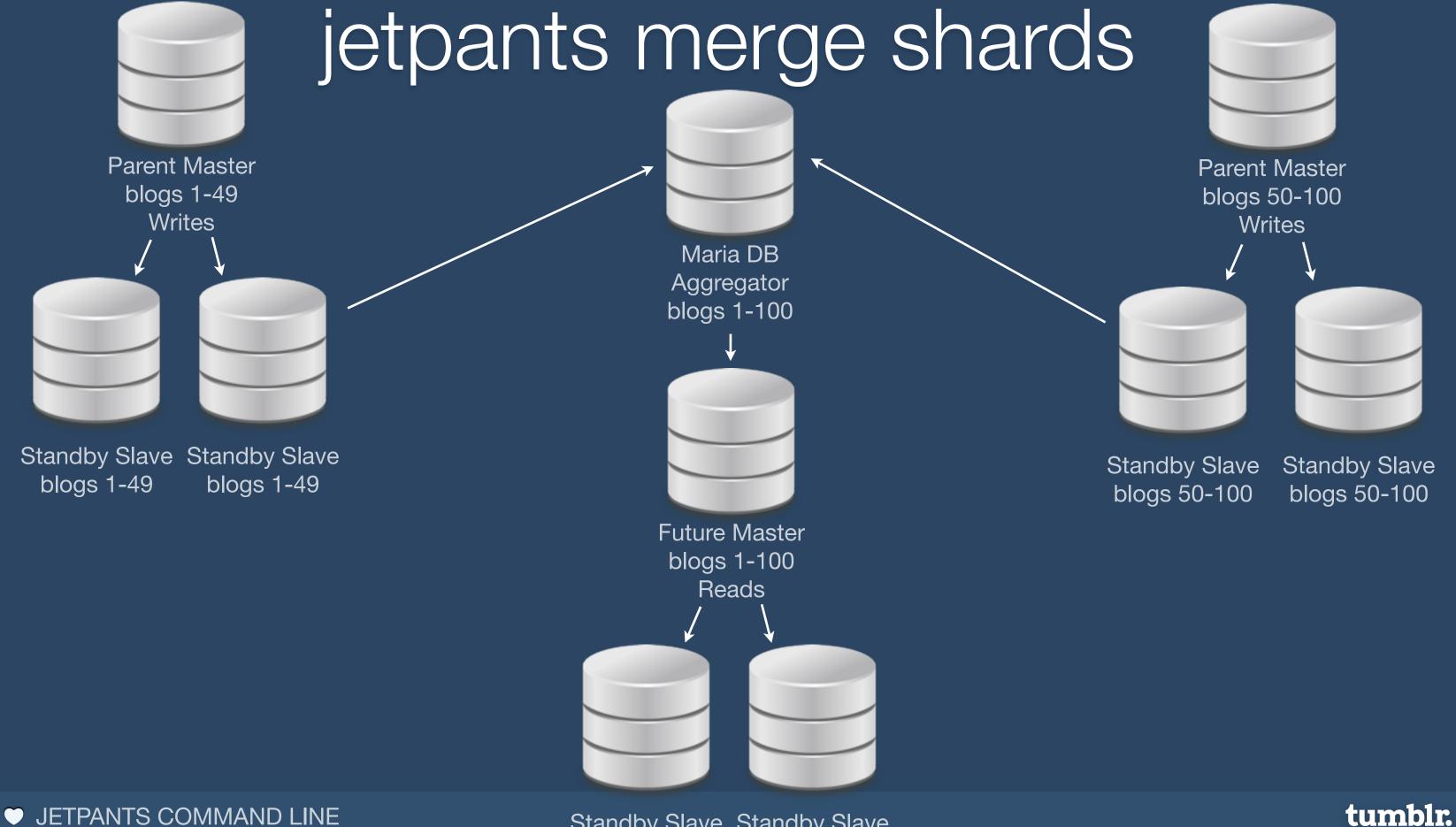
- Mariadb as an aggregator node (for its multisource replication)
- → Export data from each shard
- Import data into aggregator node and new master at same time
- → Clone new slaves from new master
- Pt-query-checksum during testing and a validator for each live run

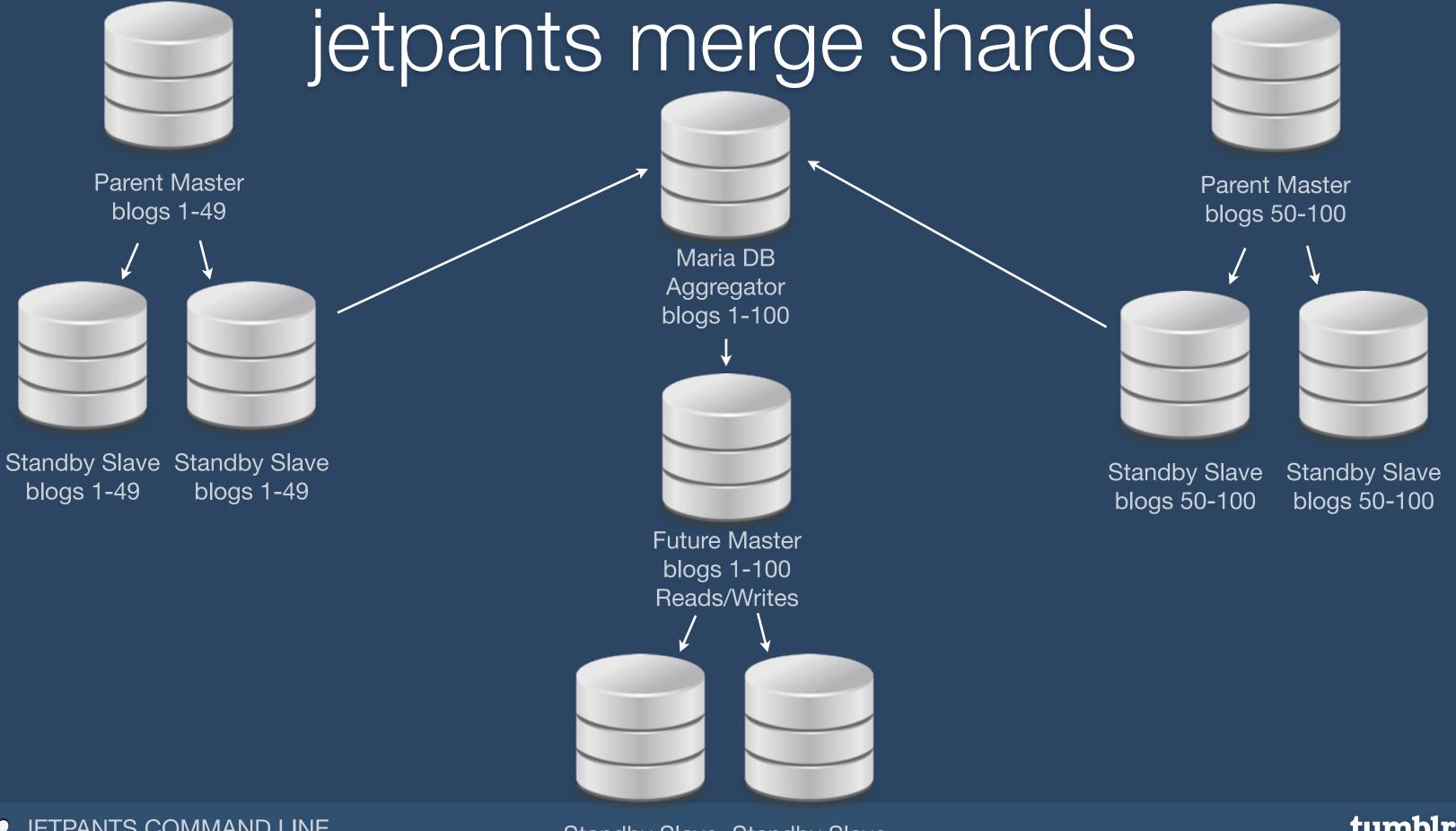














jetpants alter table

jetpants alter table

- → Wraps Percona's pt-online-schema-change
- Tracks alters via jetpants
- → Does live alter table on master
- → No downtime
- → Shard command to do the alter on all shards

jetpants capacity plan

jetpants capacity plan

- Tracks data usage on all pools
- → Calculates growth rates
- Finds outliers for abnormally high growth
- → Rollups of all your hardware usage and states
- Emails reports every morning

jetpants capacity plan

The following pools ha	ave to many	standby s	laves :					
db-tumblelogs								
Warning Mounts								
	00							
blogs-10000000-1099999	99							
Usage and Time Left The 'GB pe	r day' and lly decayin	'Days left'	fields are u	sing a q	growth rate	that is cal	ulated b	y taking
	pool name		ata Size GB p	er day	Days left	(until criti	cal)	
	postsreff		238.82	4.11	237.55			
	primary		344.42	1.91	456.27			
	-1-999999		593.63	0.56	N/A			
blogs-999999	-infinity		48.12	3.33	350.53			
Day Over Day								
	pool name	today	1 day ago 2 d	ays ago	7 days ago	14 days ago		
dbm-	postsreff	4.24	4.17	3.78				
	primary	1.69	2.07	2.29				
	-1-999999	0.53	0.58	0.58				
blogs-999999	-infinity	3.61	3.33	2.87	2.81	0.00		
New Outliers								
	days in 2 h	our blocks	to the same 2	hour b	lock 7, 14,	21, 28 days	ago	
								Danie Manka
Compare the last 3	Pool Name		Start Time			End Time	Usage	Prev Weeks
Compare the last 3	Pool Name primary	09/20/	Start Time 2013 00:02:19		09/20/2013		Usage 2.25	
Compare the last 3 (09/20/						
Compare the last 3 (primary	09/20/ abasenode			09/20/2013			
Compare the last 3 o	primary		2013 00:02:19		09/20/2013	02:04:19		0.26
Compare the last 3 of the l	primary	abasenode	2013 00:02:19	55node	09/20/2013	02:04:19 abasenode		0.26
Compare the last 3 of the l	primary	abasenode 	2013 00:02:19	55node 	09/20/2013	abasenode		total 253
Hardware status status allocated spare	primary	abasenode 0 0	2013 00:02:19	55node 247 6	09/20/2013	abasenode 6		total 253 8 0
Hardware status status allocated spare claimed	primary	abasenode 0 0 0	2013 00:02:19	55node 247 6 0	09/20/2013	abasenode 6 2		total 253



Q&A