



# Practicing DBA's Guide to the PBXT Storage Engine

## FOSDEM 2009

Vladimir Kolesnikov

PrimeBase Technologies GmbH

[www.primebase.org](http://www.primebase.org)



# Contents

- PBXT Features
- Engine Design & Architecture
- Installation
- Disk Performance
- Monitoring and Tune-up
- Backup
- Future work



# PBXT Engine Features

- Open Source (GPL v.2)
- A pluggable engine for MySQL 5.1, MySQL 6.x and Drizzle
- Transactional (BEGIN, COMMIT, ROLLBACK)
- ACID-compliant
- Row-level locks
- Referential integrity (FKs incl. NO ACTION)
- Fast rollback and recovery

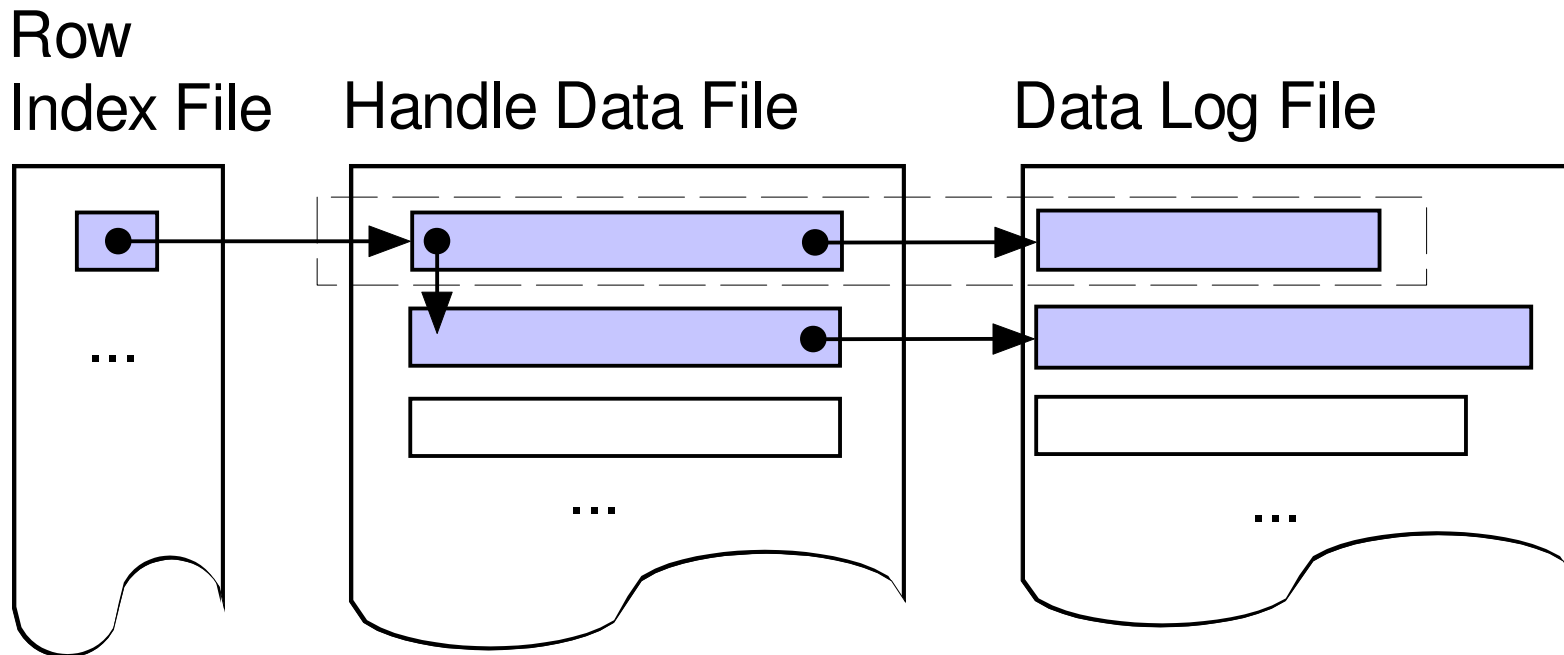


# PBXT Engine Design

- Write-once (log-based)
- Disk-based MVCC
- Fast, minimalistic commits
- No in-place updates
- No undo, fast recovery
- File-per-table



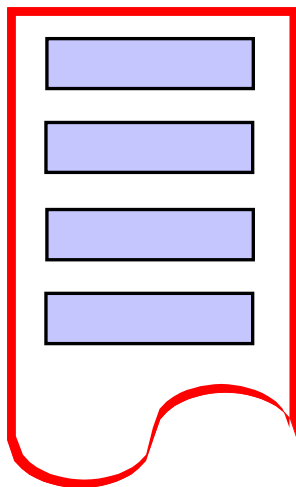
# PBXT Table Data Layout





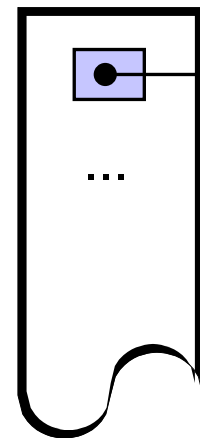
# INSERT, UPDATE, DELETE

Transaction  
Log File

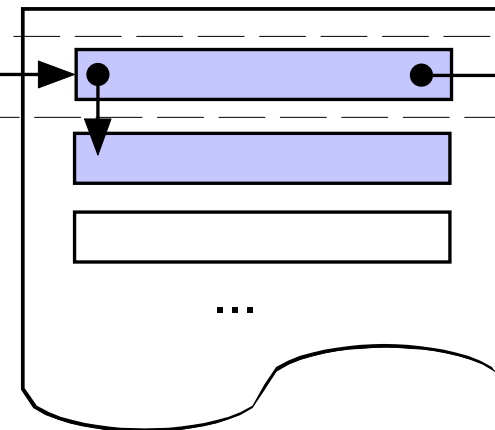


Row

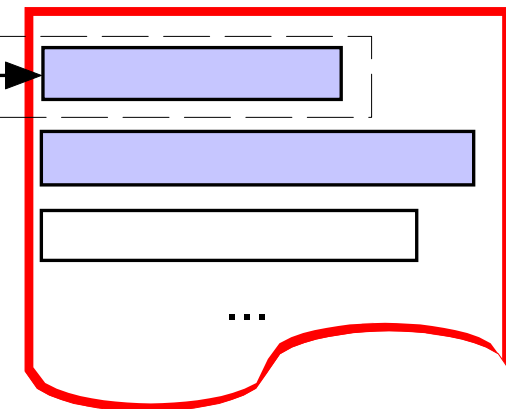
Index File



Handle Data File



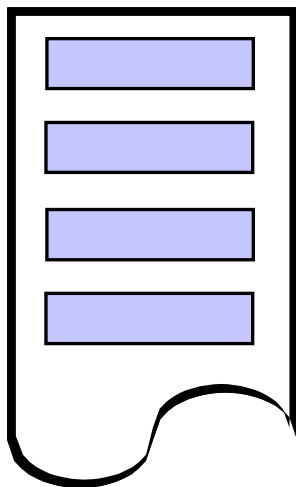
Data Log File





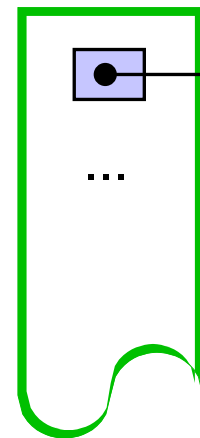
# SELECT

Transaction  
Log File

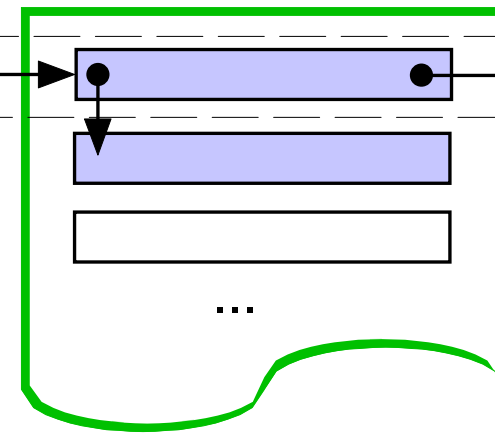


Row

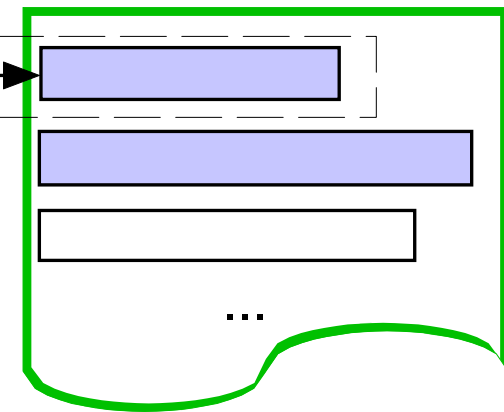
Index File



Handle Data File



Data Log File



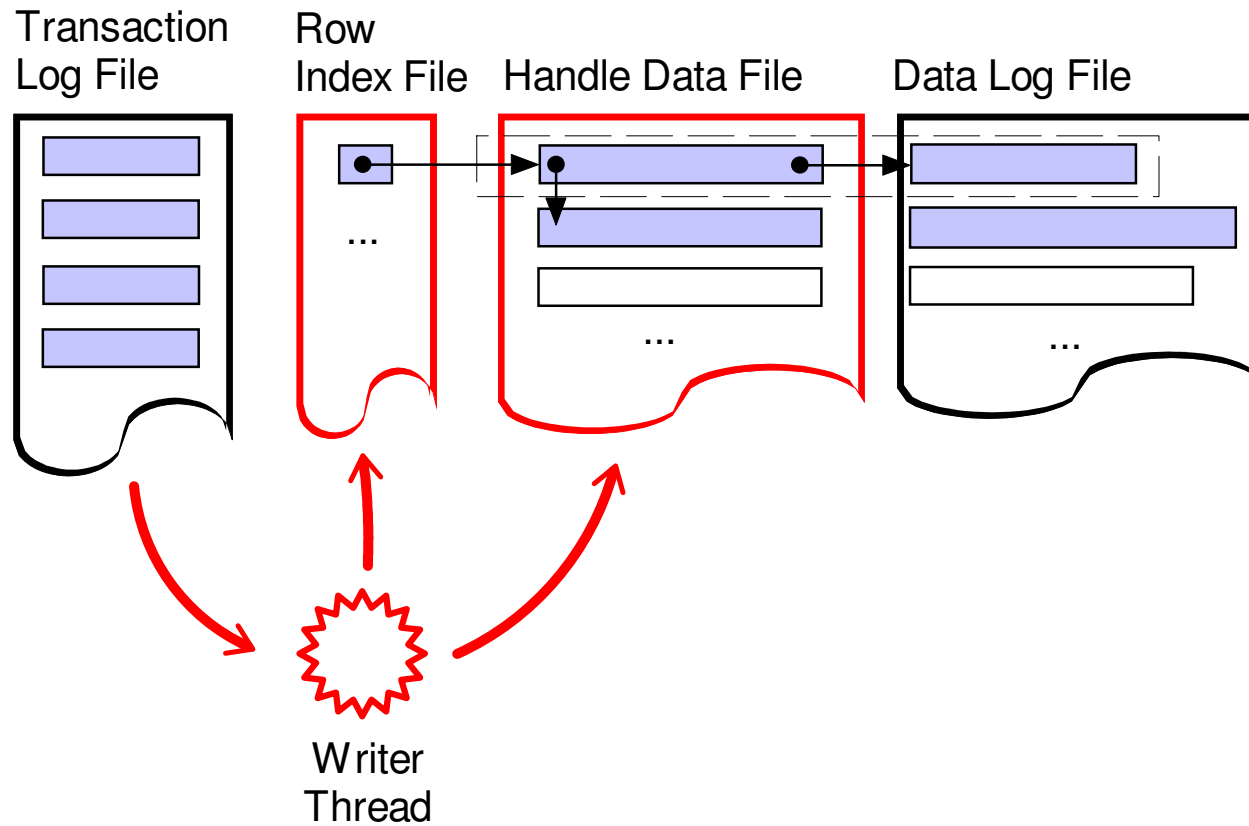


# PBXT Background Threads

- Writer
- Sweeper
- Checkpointer
- Compactor

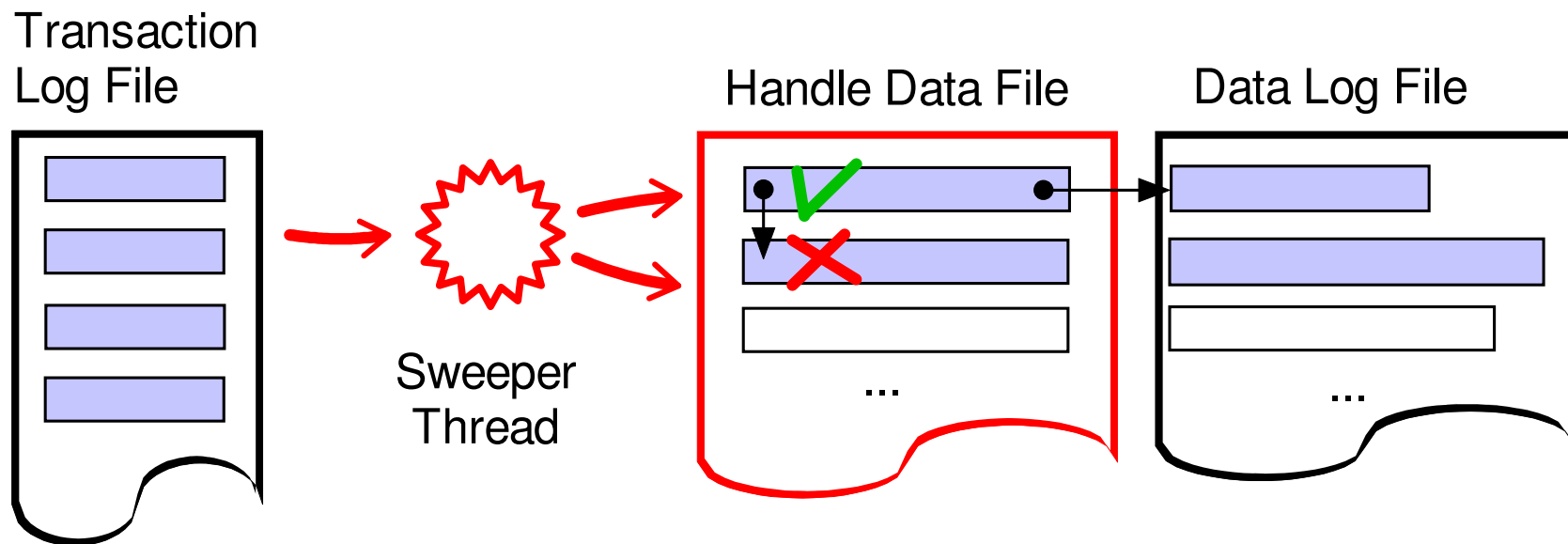


# The Writer Thread



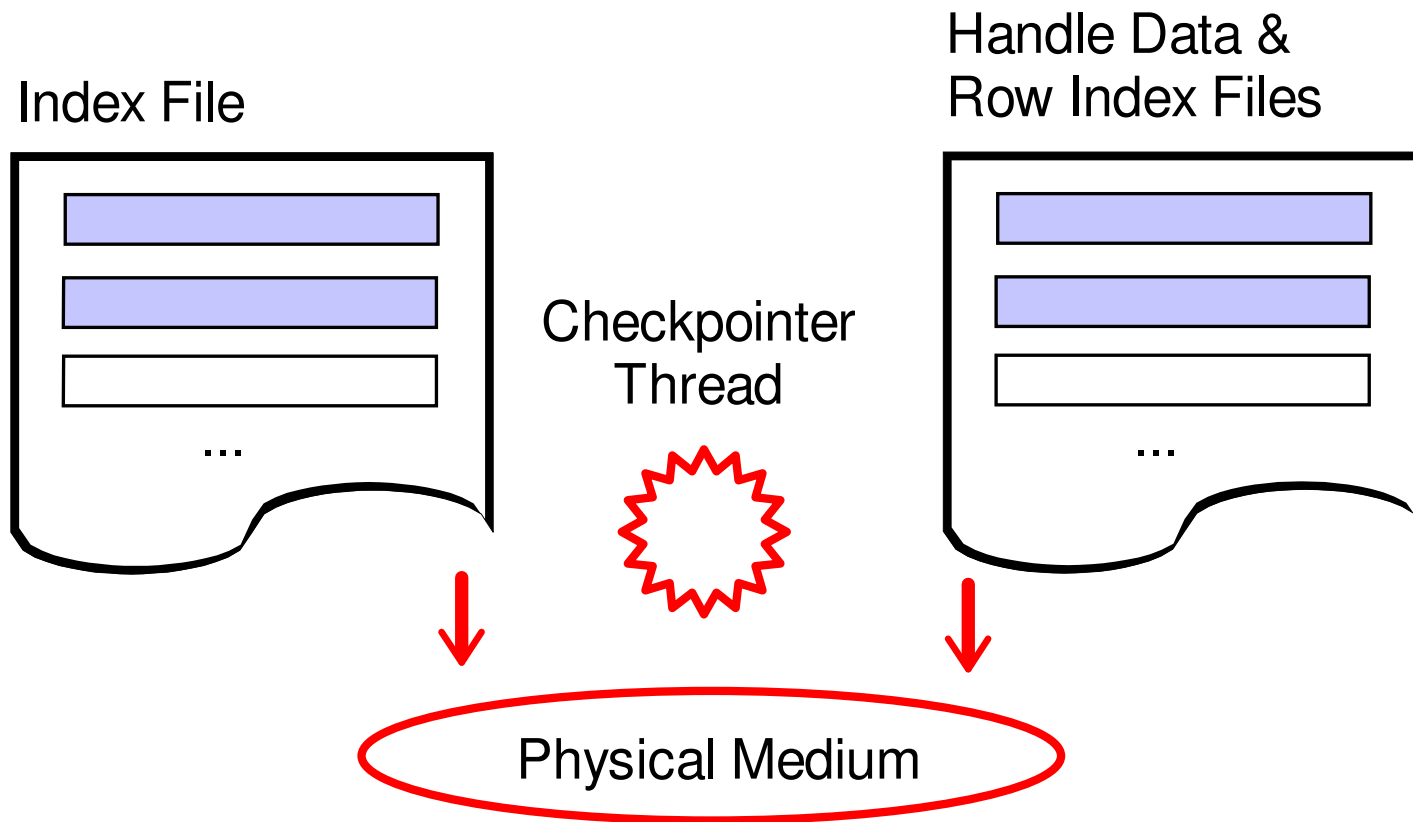


# The Sweeper Thread



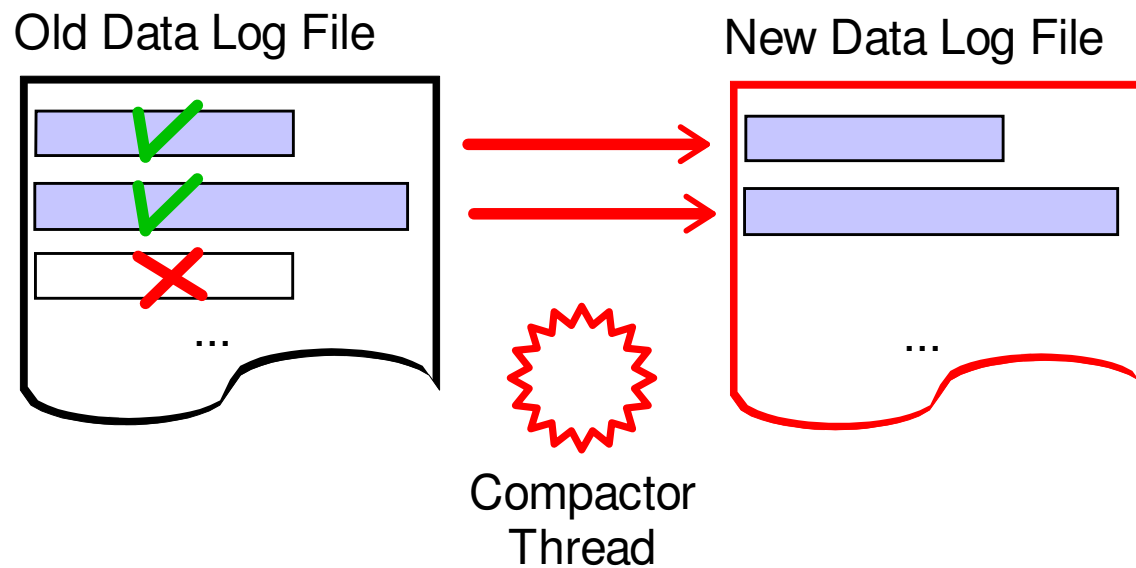


# The Checkpointer Thread





# The Compactor Thread





# PBXT Installation

- Open Source (GPL v.2)
- Hosted at Launchpad (<http://launchpad.net/pbxt>)
- Available as binary for MySQL 5.1, MySQL 6.0, Drizzle, included into XAMPP
- Tested on Linux, Solaris (x86 and SPARC), BSD, OSX, Win32/64

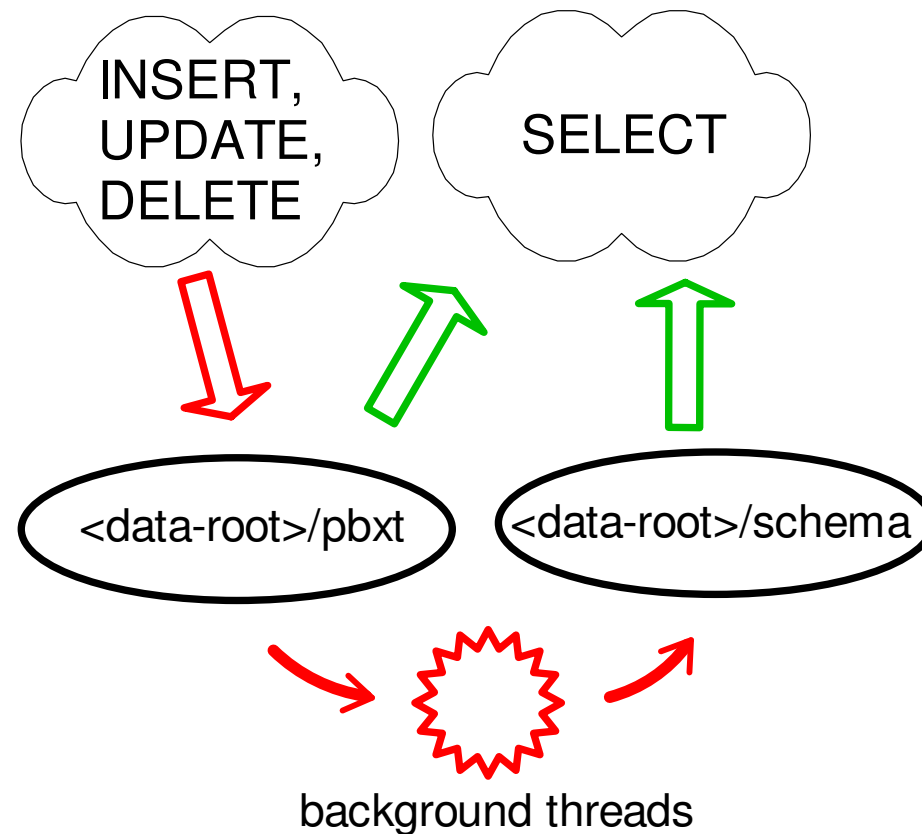


# PBXT Database Deployment General

- Foreground write operations are always sequential
- Foreground write operations modify only the <data-root>/pbxt/ directory
- Transaction log is never read in foreground (except startup recovery)



# PBXT Deployment on HDD



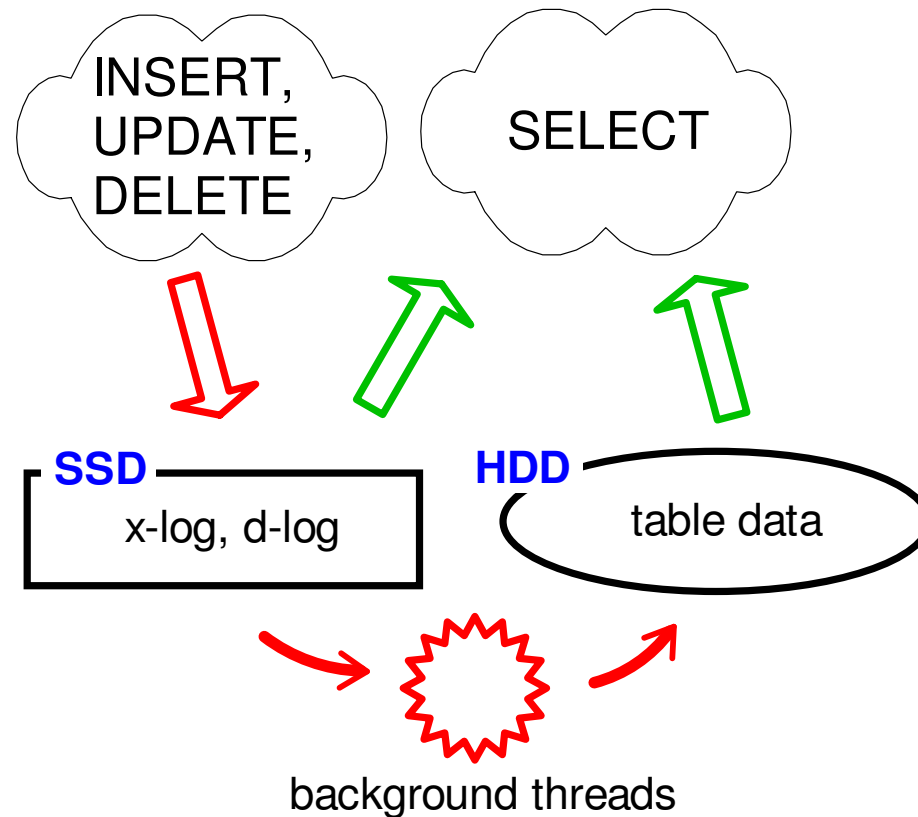


# Optimizing PBXT for SSD

	SSD	HD
Random writes	269/s	175/s
Random writes (in cache)	271/s	427,204/s
Sequential writes	112,961/s	71,975/s
Sequential writes (in cache)	559,003/s	721,709/s
Random reads	6,925/s	225/s
Random reads (from cache)	650,533/s	676,956/s



# Combining HDD & SSD





# PBXT Monitoring and Tune-up

- PBXT MySQL Variables
- Table Options
- SHOW ENGINE STATUS
- I\_S.PBXT\_STATISTICS Table & xtstat



# PBXT MySQL Variables

- Cache control
  - pbxt\_index\_cache\_size
  - pbxt\_record\_cache\_size
  - pbxt\_log\_cache\_size



# PBXT MySQL Variables

- Transaction Manager
  - pbxt\_log\_file\_threshold
  - pbxt\_transaction\_buffer\_size
  - pbxt\_checkpoint\_frequency
  - pbxt\_sweeper\_priority
  - pbxt\_auto\_increment\_mode



# PBXT MySQL Variables

- Data Log
  - pbxt\_data\_log\_threshold
  - pbxt\_garbage\_threshold
  - pbxt\_log\_buffer\_size
  - pbxt\_log\_file\_count
  - pbxt\_offline\_log\_function



# PBXT MySQL Variables

- Storage Allocation
  - `pbxt_data_file_grow_size`
  - `pbxt_row_file_grow_size`



# Table Options

ENGINE [=] PBXT

| AUTO\_INCREMENT [=] value

| [DEFAULT] CHARACTER SET [=] charset\_name

| [DEFAULT] COLLATE [=] collation\_name

| COMMENT [=] 'string'

| AVG\_ROW\_LENGTH [=] byte\_length



# SHOW ENGINE STATUS

```
mysql> show engine pbxt status\G
***** 1. row *****
Type: PBXT
Name:
Status:
090203 19:07:59 PBXT 1.0.08 RC STATUS OUTPUT
Record cache usage: 65651
Record cache size: 8388608
Record cache high: 65651
Index cache usage: 16384
Index cache size: 33554432
Log cache usage: 295128
Log cache size: 16756712
Data log files:
```



# PBXT\_STATISTICS Table

- Total 48 counters
- Transaction Statistics
- SQL Statistics
- Table and Index I/O Counters
- Table and Index Cache Status & Counters
- D-Log and X-Log Counters



# The XTSTAT Tool

- IOSTAT-like instant monitoring interface
- Local/remote monitoring
- Output filtering
- Suitable for scripting and logging



# PBXT Backup

- On-line Backup with  
`BEGIN/SELECT * /COMMIT`
- MVCC prevents phantom reads



# PBXT – What's next?

- Semi-durable Tables
- Memory Tables (with durability option)
- MySQL 6.0 Optimizations
- Flexible Data and Index Compression



# Thank you!

## Q&A

<http://www.primebase.org>

<https://launchpad.net/pbxt>

<http://pbxt.blogspot.com>