



FOSDEM 2026

LSM tree vs. B-Tree

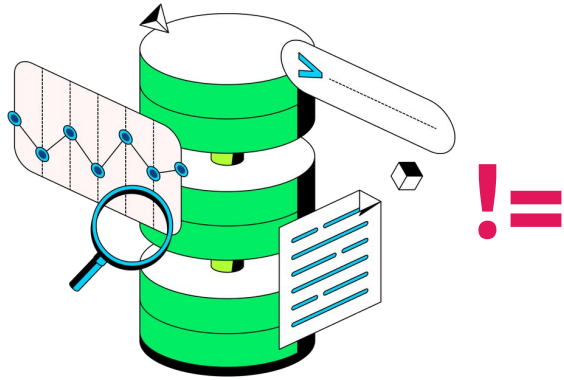
RocksDB and WiredTiger

for Cloud-Native Distributed Databases



Franck Pachot
Developer Advocate
MongoDB

LSM tree for write
B-tree for reads



Write-heavy
workloads
are also
read-intensive

- check if the key exists (duplicate key)
- can't upsert secondary indexes
- check lock flags
- check MVCC visibility
- check foreign key parent



Databases need sorted structures

Creating data (upsert)
can write to a log or heap table

but: **Queries** must find data
efficiently with point or range queries

Then data must be indexed
with a sorted structure

With lots of data, index doesn't fit in RAM

Is this sorted?



Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean
Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Anaïs
Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami
Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima
Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid
Fiona Florin Franck Gabriel Goran Greta Guillaume Haruto Hassan
Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou
Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier
Clara Cornelius Cyrus Dae Dana Daniela Darwin Dario Dawit Dimitri
Marcel Maria Mateo Miguel Nadia
Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz
Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina
Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara

Is this sorted?



Needs only a few additional pointers

Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean #211
Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Anaïs #211
Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami #212
Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima #206
Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid #207
Fiona Florin Franck Gabriel Goran Greta Guillaume Haruto Hassan #208
Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou #210
Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier #205
Clara Cornelius Cyrus Dae Dana Daniela Darwin Dario Dawit Dimitri #204
Marcel Maria Mateo Miguel Nadia #203
Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz #213
Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina #201
Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara #209

Is this sorted?



Physically sorted blocks logically ordered with pointers

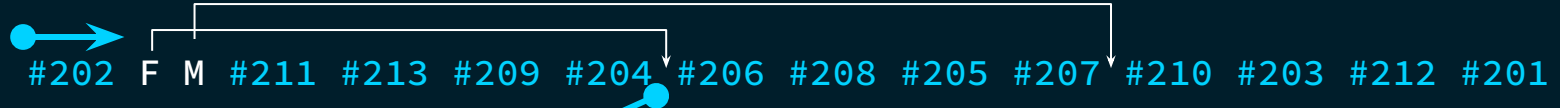
#202 #211 #213 #209 #204 #206 #208 #205 #207 #210 #203 #212 #201

#201	Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean	
#202	Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Anaïs	#211
#203	Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami	#212
#204	Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima	#206
#205	Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid	#207
#206	Hona Florin Franck Gabriel Goran Greta Guillaume Haruto Hassan	#208
#207	Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou	#210
#208	Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier	#205
#209	Clara Cornelius Cyrus Dae Dana Daniela Darwin Dario Dawit Dimitri	#204
#210	Marcel Maria Mateo Miguel Nadia	#203
#211	Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz	#213
#212	Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina	#201
#213	Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara	#209

Is this sorted?



Skip list is a link list with shortcuts (fast-forward pointers)



#201	Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean	
#202	Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Anaïs	#211
#203	Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami	#212
#204	Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima	#206
#205	Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid	#207
#206	Fiona Florin Franck Gabriel Goran Greta Guillaume Haruto Hassan	#208
#207	Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou	#210
#208	Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier	#205
#209	Clara Cornelius Cyrus Dae Dana Daniela Darwin Dario Dawit Dimitri	#204
#210	Marcel Maria Mateo Miguel Nadia	#203
#211	Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz	#213
#212	Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina	#201
#213	Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara	#209



B-Tree

B+tree (levels)



Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean
Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Anaïs
Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami
Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima
Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid
Fiona Florin Franck Gabriel Goran Greta Guillaume Haruto Hassan
Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou
Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier
Clara Cornelius Cyrus Dae Dana Daniela Darwin Dario Dawit Dimitri
Marcel Maria Mateo Miguel Nadia
Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz
Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina
Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara

B+tree (seek)



#001 Aa #101 H #102 Ng #103

#101 Aa #202 And #211 Ben #213 Cl #208 Din #204 Fi #206

#102 H #208 Ji #205 Ko #207 Mar #210

#103 Ng #203 Raf #212 S #201

#201 Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean

#202 Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Anaïs #211

#203 Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami #212

#204 Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima #206

#205 Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid #207

#206 Fiona Florin **Franck** Gabriel Goran Greta Guillaume Haruto Hassan #208

#207 Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou #210

#208 Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier #205

#209 Clara Cornelius Cyrus Dae Dana Daniela Darwin Dario Dawit Dimitri #204

#210 Marcel Maria Mateo Miguel Nadia #203

#211 Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz #213

#212 Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina #201

#213 Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara #209

B+tree (WiredTiger)

#001 Aa #101 ← H #102 Ng #103

#101 Aa #202 And #211 Ben #213 Cl #209 Din #204 Fi #206

#102 H #208 Ji #205 Ko #207 Mar #210

#103 Ng #207 Raf #212 S #201

#201 Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean

#202 Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Annis

#203 Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami

#204 Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima

#205 Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid

#206 Fiona Florin **Franck** Gabriel Garon Grata Guillaume Haruto Hassan

#207 Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou

#208 Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier

#209 Clara Cornelius Cyrus Dae Dana Daniela Darwin Dario Dawit Dimitri

#210 Marcel Maria Mateo Miguel Nadia

#211 Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz

#212 Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina

#213 Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara

B+tree (split)



#001 Aa #101 H #102 Ng #103

#101 Aa #202 And #211 Ben #213 Cl #209 Din #204 Fi #206

#102 H #208 Ji #205 Ko #207 Mar #210

#103 Ng #203 Raf #212 S #201

#201 Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean

#202 Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Anaïs #211

#203 Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami #212

#204 Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima #206

#205 Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid #207

#206 Fiona Florin Franck Gabriel Goran Greta Guillaume Haruto Hassan #208

#207 Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou #210

#208 Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier #205

#209 Clara Cornelius Cyrus Dae Dina Daniela Darwin Dario Dawit Dimitri #204

#210 Marcel Maria Mateo Miguel Nidia #203

#211 Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz #213

#212 Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina #201

#213 Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara #209

B+tree (split)



#001 Aa #101 H #102 Ng #103

#101 Aa #202 And #211 Ben #213 Cl #209 Din #204 Fi #206 #213 Go

#102 H #208 Ji #205 Ko #207 Mar #210

#103 Ng #203 Raf #212 S #201

#201 Sabrina Said Salim Samir Sandra Santiago Sara Sasha Savita Sean

#202 Aarav Abdul Adelaide Adriana Ahmed Akira Alessandro Amara Amir Araïs #211

#203 Nguyen Noor Nour Oscar Otto Oscar Otto Pablo Priya Qi Qin Rami #212

#204 Dinesh Ekaterina Elena Emeka Emiko Enrique Esme Fatima #206

#205 Jin Juan Judith Kaito Kamala Karim Katya Keisha Kenji Khalid #207

#206 Fiona Florin Franck Gabriel #213

#207 Koen Laila Lakshmi Lars Leila Leonid Liang Lucia Luis Mamadou #210

#208 Helena Hiroshi Ibrahim Ignacio Ingrid Ismael Ivana Jamal Javier #205

#209 Clara Cornelius Cyrus Dae Dina Daniela Darwin Dario Dawit Dimitri #204

#210 Marcel Maria Mateo Miguel Nidia #203

#211 Anders Anika Antonio Aria Arjun Asha Aurelia Beatriz #213

#212 Rafael Rajesh Rania Rashid Regina Rei Renata Ricardo Riko Rina #201

#213 Benedikt Bianca Branko Brigitte Carlos Carmen Chandra Chiara #209

#213 Goran Greta Guillaume Haruto Hassan #208



LSM tree

LSM tree (read)



MemTable

Beatriz Hiroshi Lucia Said Salim

SST01

Adelaide Aar Abdul Asha Ben Brigitte Carmen
Enrique Goran Hui Haruo Hsueh
Liang Lucia Marta Miguel Nadia Naoki Nour Otto
Said Santiago Savita Sasha

SST02

Aara Aki Amandis Amir Aurelia Branko Carlos Daphne Dario Ekaterina Florin
Gabriel Greta Ignacio Ivan Jin Kamal Katherine John Lakshmi Laila Luis
Maria Mateo Miriam Nawal Oscar Pablo Priya Pooja Rania Rajesh Regina Sean

SST03

Anders Bianca Brigitte Cyrus Darwin Fiona Gabriel Gisele Ismael Ivan
Judith Kaito Leonid Minar Nour Pablo Renata Rashid Salvador Samir Sandra
Sara Sebastian

bloom filters
min/max values
sort-merge

LSM tree (write)



MemTable

Beatriz Hiroshi Lucia Said Salim

insert, update (upsert) , delete (tombstone)

flushed to a new sorted string table (SST) file when full

SST04 ...

SST01

Adelaide Aaron Abdul Asha Ben Brigitte Carmen Chiara Clara Darwin Davit
Enrique Goran Hadi Haruto Helena Ingrid Kaito Kamala Keisha Lars Leila
Liang Lucia Marla Miguel Nadia Naoki Nour Otto Qi Rafael Rani Renata Riko
Said Santiago Savita Sasha

SST02

Aarav Abdul Anaïs Amir Aurelia Branko Carlos Daphne Dario Ekaterina Florin
Gabriel Greta Ignacio Ivan Jin Kamal Katherine Koen Lakshmi Laila Luis
Maria Mateo Miriam Nawal Oscar Pablo Priya Rami Rania Rajesh Regina Sean

SST03

Anders Bianca Brigitte Cyrus Darwin Fiona Gabriel Gisele Ismael Ivan
Judith Kaito Leonid Minar Nour Pablo Renata Rashid Salvador Samir Sandra
Sara Sebastian

LSM tree (compaction)

SST02

Aarav Abdul Anaïs
Gabriel Greta Ignacio
Maria Mateo Miriam

Compaction:

chooses sorted runs to **sort-merge** into a new SST file:

- reduces the space amplification (MVCC versions, tombstones)
- reduces read amplification (number of files to merge on read)

Florin
Luis
Sean

SST03

Anders Bianca Brigt
Judith Kaito Leonid
Sara Sebastian

an
Sandra

SST05

Aarav Abdul Anaïs Amir Anders Aurelia Bianca Branko Brigitte Carlos Cyrus
Daphne Dario Darwin Ekaterina Fiona Florin Gabriel Gisele Greta Ignacio
Ismael Ivan Jin Judith Kaito Kamal Katherine Koen Lakshmi Laila Leonid
Luis Maria Mateo Miriam Minar Nawal Nour Oscar Pablo Priya Rami Rania
Rajesh Rashid Regina Renata Salvador Samir Sandra Sara Sean Sebastian



B-Tree

one value has one place
maintained sorted on write
faster on read

Low read amplification when branches and frequently accessed ranges fit in memory
more RAM, esp. for internal nodes

Space amplification with random values
(25% free on average)
online rebuild, 90-10 split, avoid UUIDv4

High write amplification on block split,
and read-modify-write whole block
copy-on-write, flexible block size, reduce
fragmentation, lock-free algorithms

LSM tree

one value has multiple places
merge-sort on read (and compaction)
faster on write

High read amplification: must read
memtable, level 0 and some higher levels
bloom filters, min/max, only level 0

High space amplification with immutable
files, 2x during compaction
sharding, more full compaction

Low write amplification in memtable
but high for compaction
garbage collection, throttle writes,
combine compaction with MVCC, TTL



WiredTiger's B-tree in MongoDB

- MVCC / copy-on-write updates (no in-place changes)
- No linked leaves for range-scans (latches only during tree descent)
- Parent/child pointers (not file offsets) (fast upward/downward navigation)
- Lock-free hazard pointers (protect pages in use from eviction)
- Checksum in the address reference
- Compressed on disk (prefix, block)

MongoDB stores documents in a B+tree with monotonically increasing record ID (good time clustering, like LSM tree)

RocksDB's LSM tree in YugabyteDB

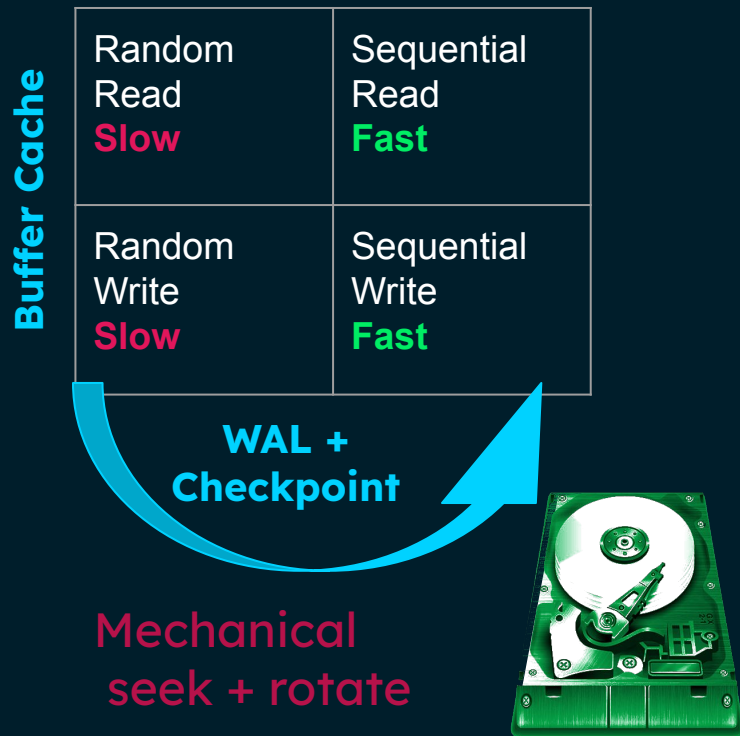
- Support only SSD (fast reads)
- Bloom filters for point queries
- Bloom on partial key for compound index
- min/max of each column for each SST file
- limit caching for full scans (block cache)

LSM tree well suited for tiered storage (cold data to S3), snapshot (easy incremental backup), branching, disaggregated storage for serverless

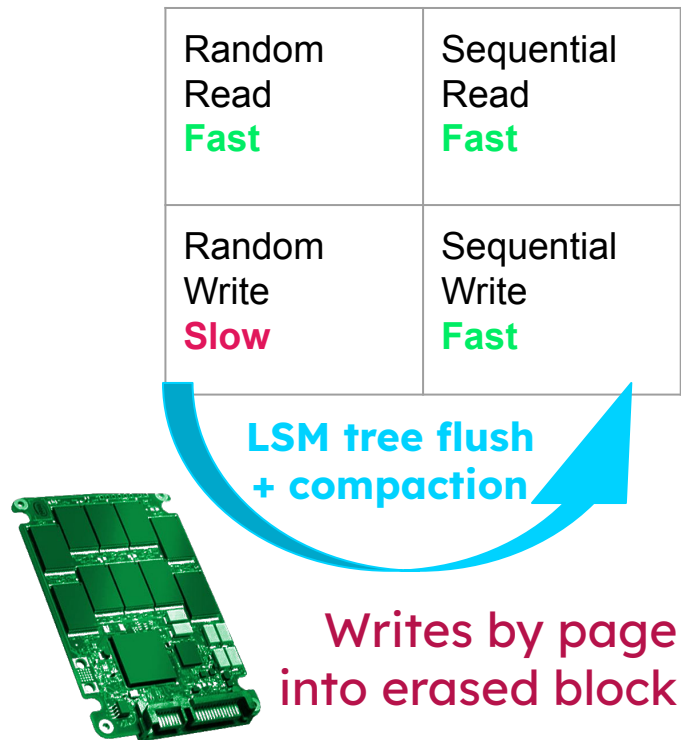
Better than B-tree for random inserts, or when less memory is available, upserts without secondary indexes

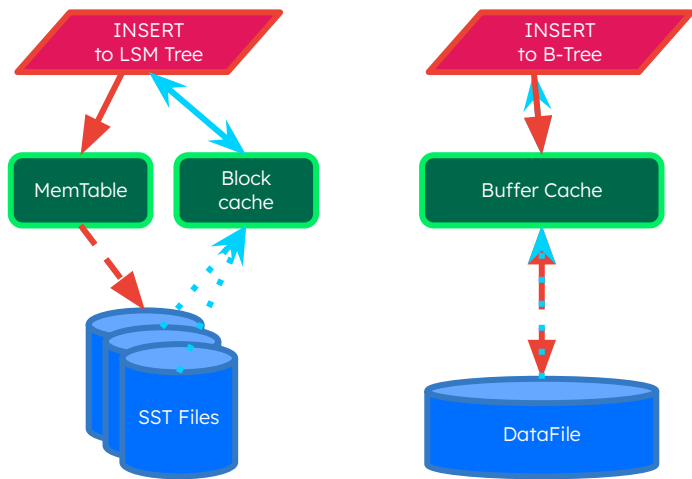


Hard Disk Drive



Solid State Drive





Write intensive?

LSM tree are faster for writes:

- because they write to memory
- B-Tree must read-modify-write

Do you have write intensive workloads?

Fast-ingest, Big Data, NoSQL, upsert in Cassandra

OLTP, SQL and document databases?

Lots of read workloads

! Writes are also read-intensive

- check if the key exists (duplicate key)
- can't upsert secondary indexes
- check lock flags
- check MVCC visibility
- check foreign key parent



B-Tree

B-tree - balanced tree data structure
Invented by Boeing Research Labs (1970)

All types of databases:

Oracle Database
MySQL (InnoDB)
Microsoft SQL Server
PostgreSQL
MongoDB (WiredTiger)
Amazon DynamoDB
Amazon Aurora DSQL
SQLite
IBM Db2, Ingres, Sybase
Microsoft Access, dBase, FoxPro

LSM tree

Log-structured merge-tree
Published in 1996 (idea 1991)

Big data and distributed SQL:

Apache Cassandra, ScyllaDB
RocksDB, LevelDB
MyRocks
HBase, BigTable
Spanner, CockroachDB
YugabyteDB, TiDB
VictoriaMetrics
ClickHouse (MergeTree)
InfluxDB (TSM engine)



Franck Pachot
Developer Advocate MongoDB

Final thoughts

LSM tree faster for write

- but most databases are read-heavy

LSM tree complex to tune

- hundreds of parameters for memory allocation, compaction algorithm

B-Tree:

- simpler configuration
- (but be careful with random vs increasing keys)
- often fixed block size (then needs heap table, or overflow) but WiredTiger has variable block size

Understanding the trade-offs and implementation
is more important than finding winner