Final Presentation

Studying state-of-the-art HTAP systems

A lecture on HTAP Systems

- Satya Sai Bharath Vemula
- Rwitam Bandyopadhyay
- Shubham Pandey

PURDUE UNIVERSITY

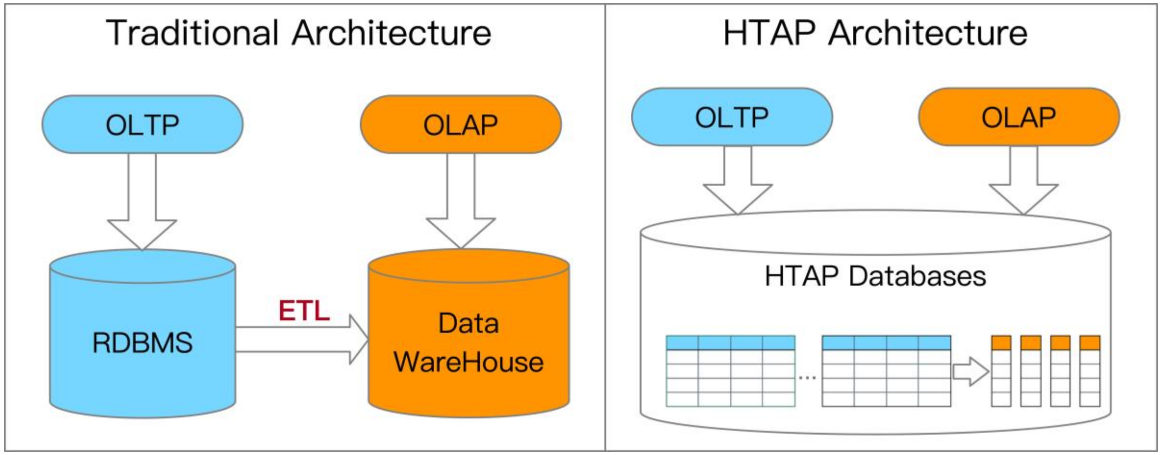
Table of Contents

1. Motivation
2. HTAP Databases
3. HTAP Techniques

Motivation

Hybrid Transactional and Analytical Processing, HTAP

- **Gartner's definition in 2014**: utilizes in-memory computing technologies to enable concurrent analytical and transaction processing on the same in-memory data store.
- **Gartner's new definition in 2018**: supports weaving analytical and transaction processing techniques together as needed to accomplish the business task.



Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Motivation

- Gartner envisioned that, HTAP techniques will be widely adopted in the business applications with real-time data analytics by 2024.
- HTAP databases have many applications in E-commerce, Finance and Banking, Fraud Detection, etc.
- For example, identify the sales trend on-the-fly in e-commerce; detecting the fraudulent transactions when processing the transactions.

Fraud Detection

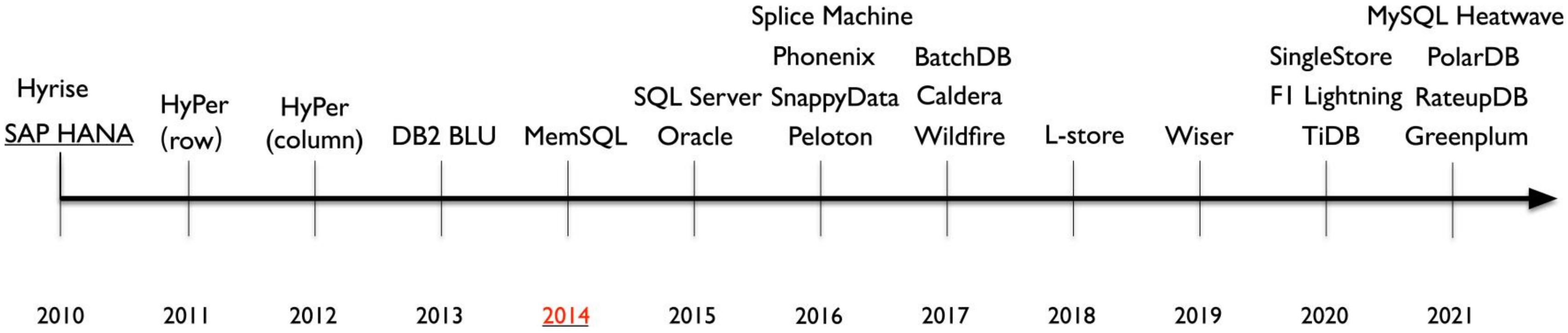
Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Motivation

Over the last decade, many **HTAP databases** have emerged

The following timeline consists of three phases:

- **Phase 1** (2010-2014): HTAP databases mainly adopt primary column store
- **Phase 2** (2014-2020): HTAP databases mainly extend the primary row store
- **Phase 3** (2020-present): HTAP databases utilize a distributed architecture

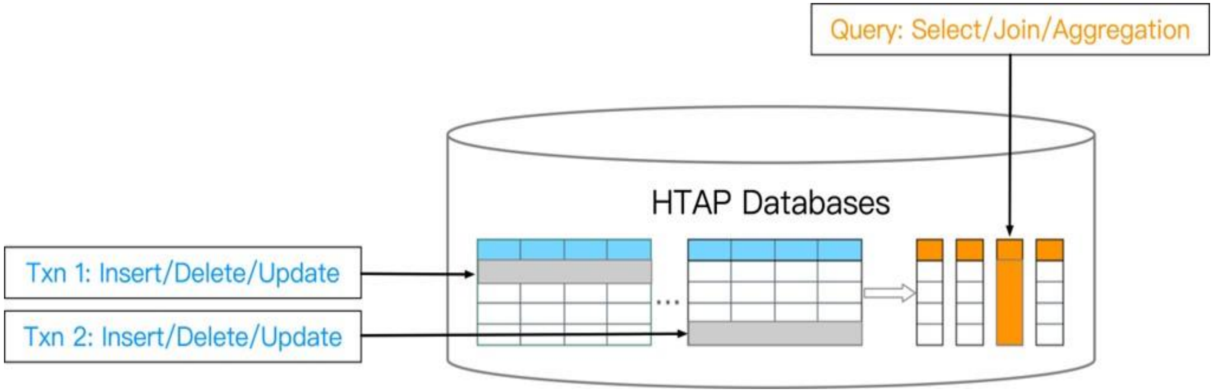


Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Motivation

- **Rule of thumb 1**: Row store is ideal for OLTP workloads
  - Row-wise, update-heavy, short-lived transactions
- **Rule of thumb 2**: Column store is best suited for OLAP workloads
  - Column-wise, read-heavy, bandwidth-intensive queries

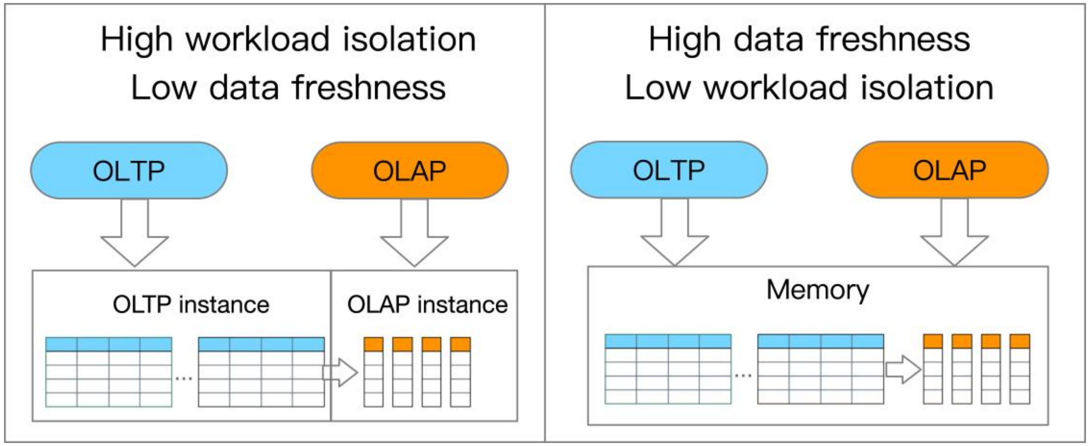
We study HTAP databases with both row store and column store


Query: Select/Join/Aggregation
HTAP Databases
Txn 1: Insert/Delete/Update
Txn 2: Insert/Delete/Update
...

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

A trade-off for HTAP databases

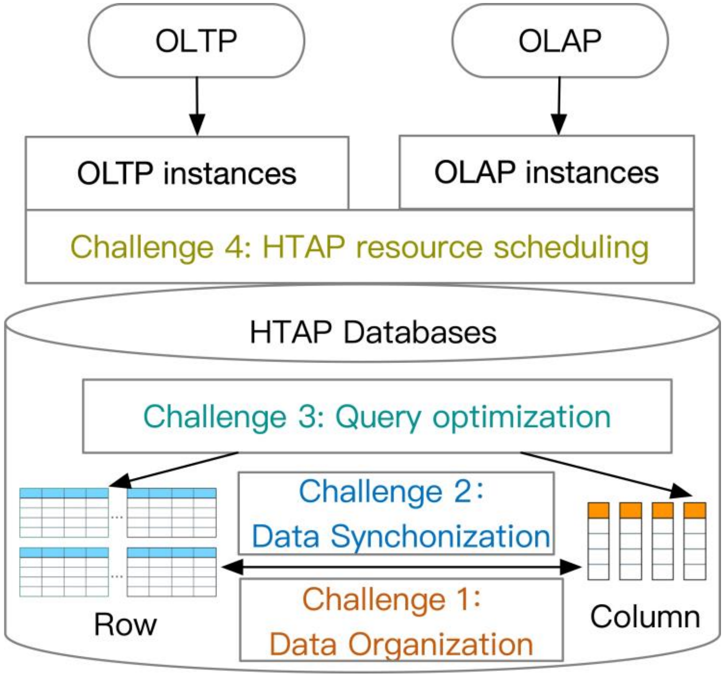
- **Workload isolation:** the isolation level of handling the mixed workloads
- **Data freshness:** the portion of latest transaction data that is read by OLAP
- Trade-off for workload isolation and data freshness
  - High workload isolation leads to low data freshness
  - Low workload isolation results in high data freshness



Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Challenges for HTAP databases

- **Challenge 1 (Data Organization)**: how to organize the data adaptively for HTAP workloads with high performance and low storage cost.
- **Challenge 2 (Data Synchronization)**: how to synchronize the data from the row store to the column store for high throughput and data freshness
- **Challenge 3 (Query Optimization)**: how to optimize the query with both row store and column store by exploring the huge plan space.
- **Challenge 4 (Resource Scheduling)**: how to schedule the resources for OLTP and OLTP instances effectively for high throughput and data freshness.



OLTP
OLAP
OLTP instances
OLAP instances
Challenge 4: HTAP resource scheduling
HTAP Databases
Challenge 3: Query optimization
Challenge 2: Data Synchronization
Challenge 1: Data Organization
Row
Column

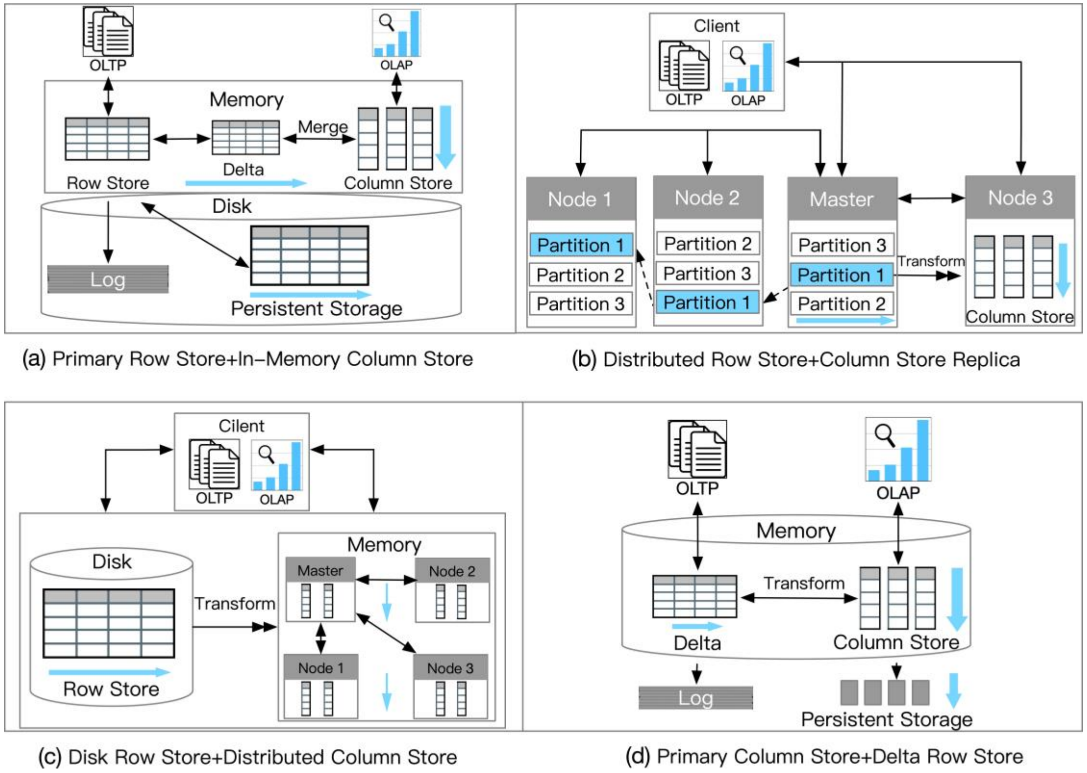
PURDUE UNIVERSITY

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

HTAP DATABASES

An Overview of HTAP Architectures

- Primary Row Store + In-Memory Column Store
- Distributed Row Store + Column Store Replica
- Disk Row Store + Distributed Column Store
- Primary Column Store + Delta Row Store



(a) Primary Row Store+In–Memory Column Store

(b) Distributed Row Store+Column Store Replica

(c) Disk Row Store+Distributed Column Store

(d) Primary Column Store+Delta Row Store

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Popular HTAP Architectures

(a) Primary Row Store+ In Memory Column Store

Pros:

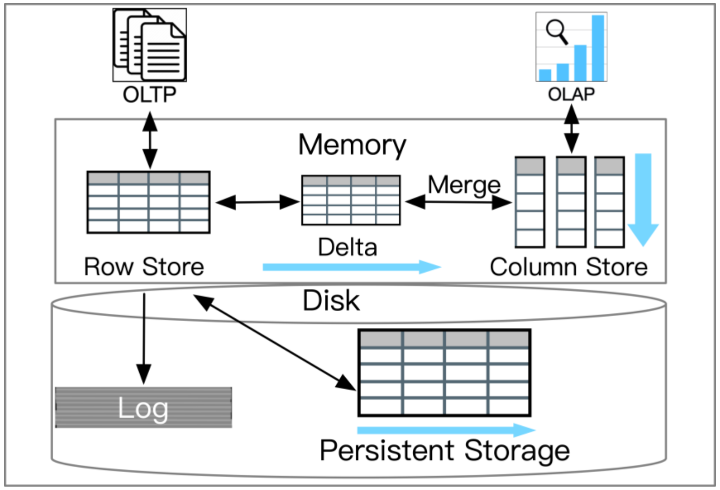
- High TP throughput,
- High AP throughput
- High data freshness

Cons:

- Low AP scalability
- Low workload isolation

Applications:

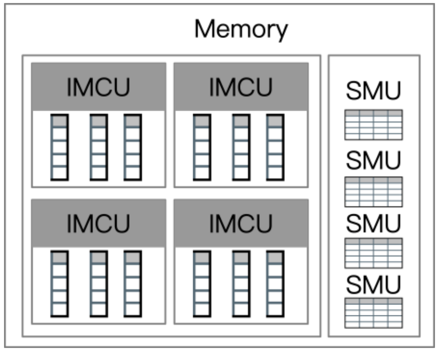
- High throughput, low scalability (e.g., banking with real-time data analytics)



(a) Primary Row Store + In–Memory Column Store

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

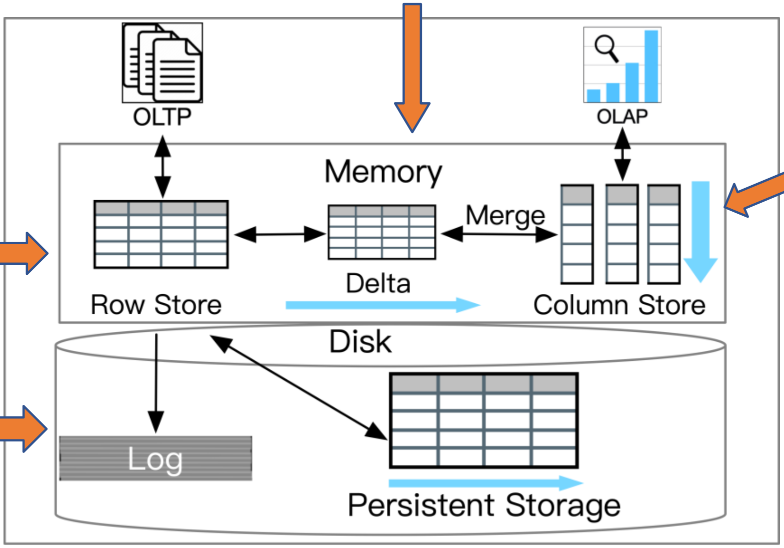
Case Study: Oracle Dual-Format



Oracle Cache

Oracle Disk-based Row Store

Oracle Snapshot Memory Unit (SMU)



(a) Primary Row Store + In–Memory Column Store

Oracle In-memory column store (IMCS)

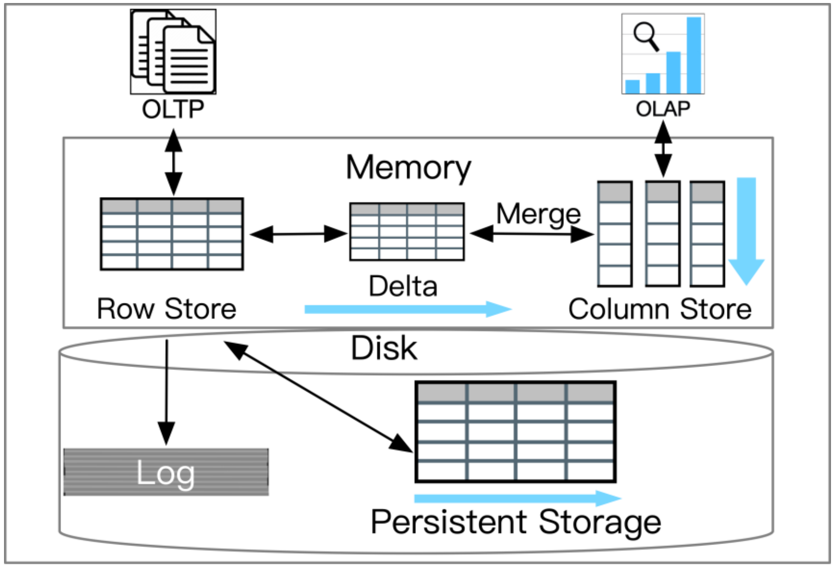
Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Challenges for HTAP Databases with architecture (a)

Problems: need to increase the AP scalability and workload isolation

Challenges: how to scale and isolate the AP (i.e., column store) while maintaining high TP & AP throughput and data freshness

Possible ways:
Scale up/out the memory capacity

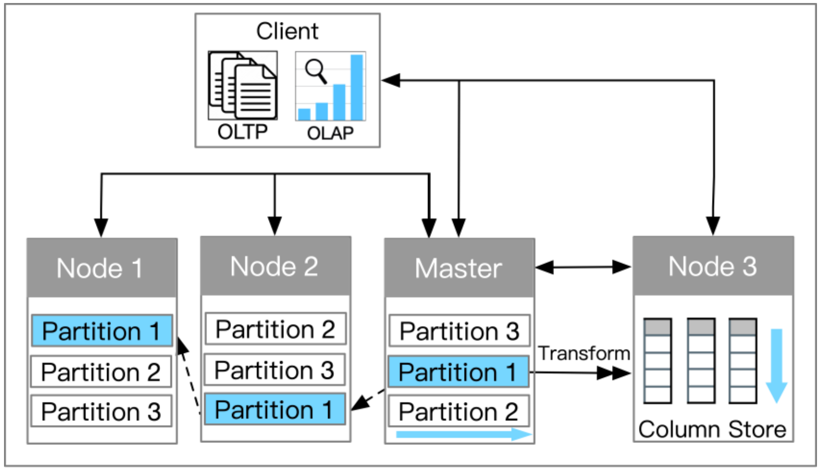


(a) Primary Row Store + In–Memory Column Store

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Popular HTAP Architectures

(b) Distributed Row Store + Column Store Replica



(b) Distributed Row Store + Column Store Replica

Pros:

- High workload isolation
- High scalability

Cons:

- Low data freshness

Applications:

- High TP & AP scalability, tolerable data freshness (e.g., E-commerce with real-time data analytics)

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

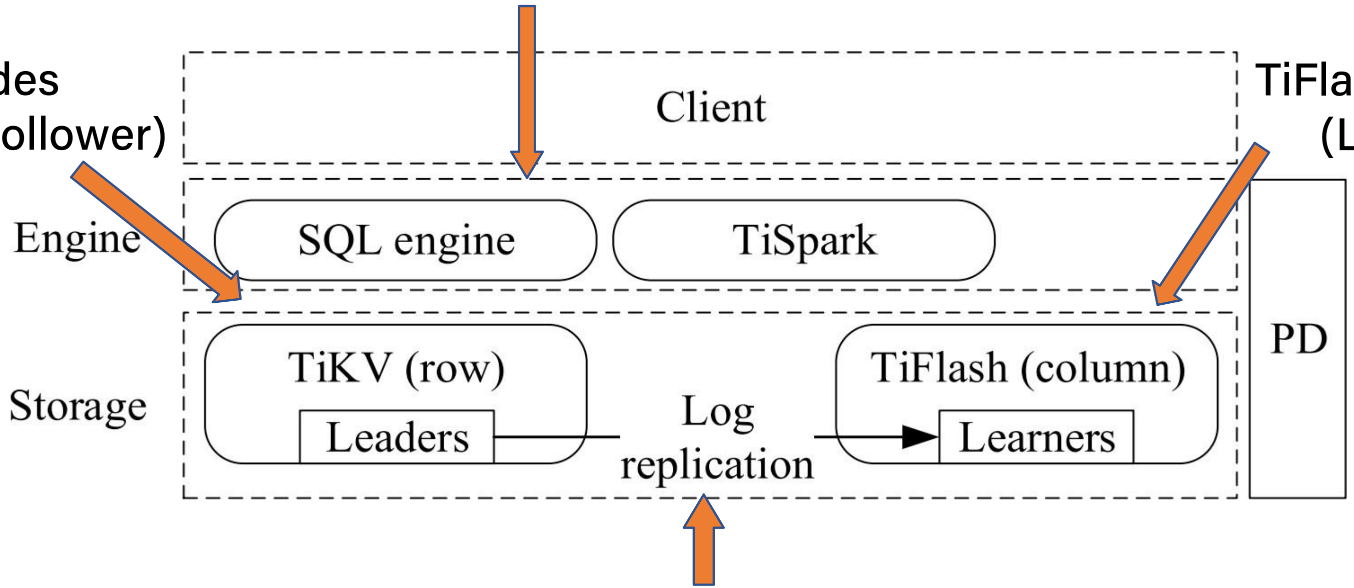
An Overview of HTAP Architectures

Case Study: TiDB

2PC+Raft for TP
Cost-based query processing for AP

TiKV nodes
(Leader and Follower)

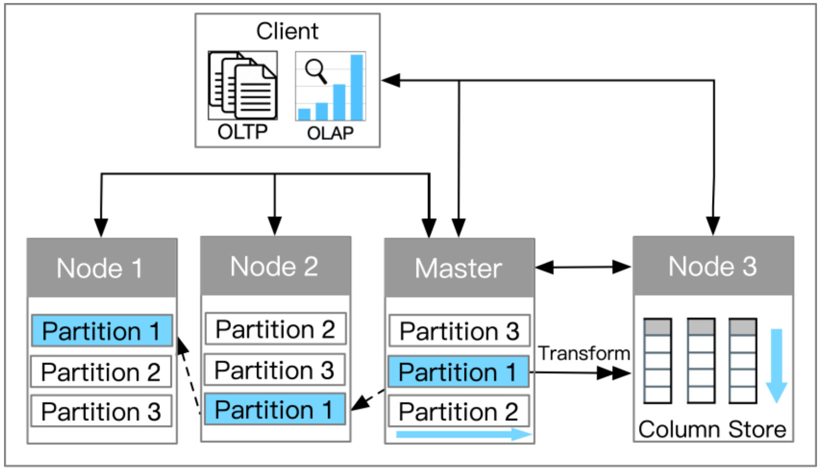
TiFlash nodes
(Learner)



Asynchronously Log-based replay+ column delta tree

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Challenges for HTAP Databases with architecture (b)



(b) Distributed Row Store + Column Store Replica

Problems: need to increase the data freshness

Challenges: how to efficiently merge the delta files to the column store

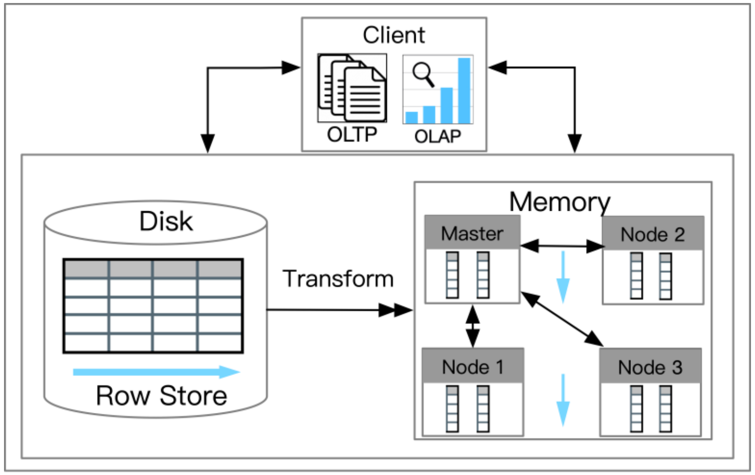
Possible solutions:

- Memory-based delta logging and shipping
- New indexing techniques for delta merging

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Popular HTAP Architectures

(c) Disk Row Store + Distributed Column Store



(c) Disk Row Store + Distributed Column Store

Pros:

- High workload isolation
- High AP throughput and scalability

Cons:

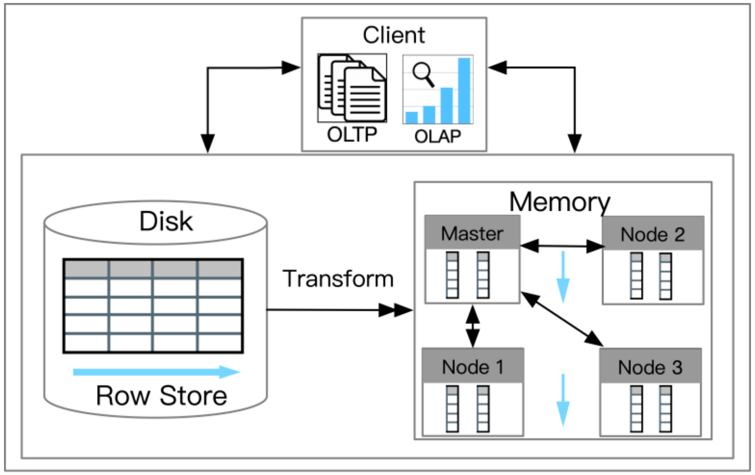
- Medium(On-premise)/Low(Cloud-based) data freshness

Applications:

- High AP scalability, tolerable data freshness (e.g., IoT applications with real-time data analytics)

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Challenges for HTAP Databases with architecture (c)



(c) Disk Row Store + Distributed Column Store

Problems: need to increase the data freshness and reduce the storage cost

Challenges: how to balance the data freshness AP throughput, and storage cost adaptively

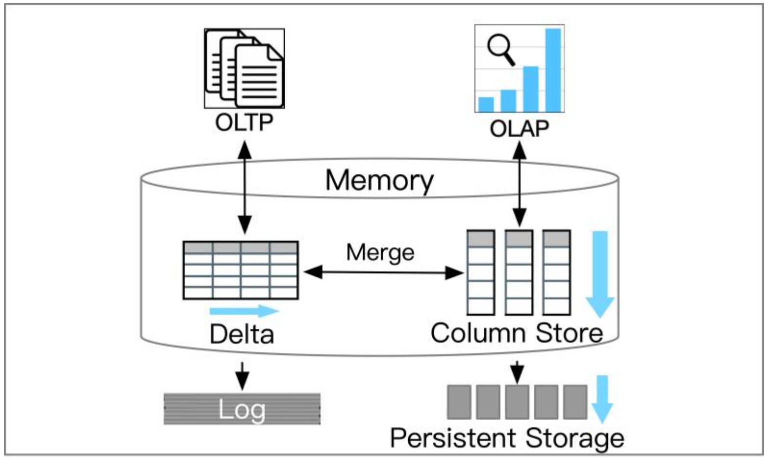
Possible solutions:

- Cost models for column data management

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Popular HTAP Architectures

(d) Primary Column Store + Delta Row Store



(d) Primary Column Store + Delta Row Store

Pros:

- High data freshness
- High AP throughput

Cons:

- Low TP scalability
- Low workload isolation

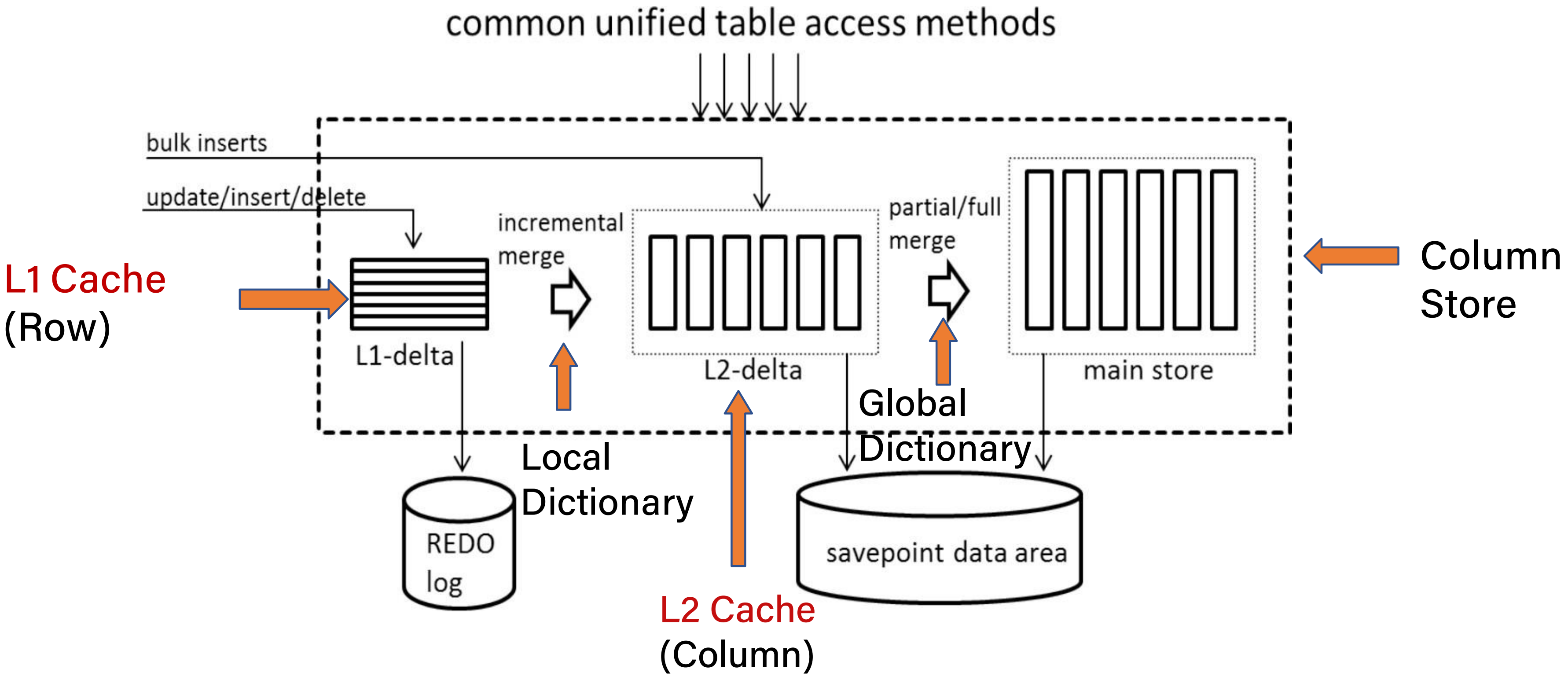
Applications:

High AP throught, High data freshness (e.g., Real-time Fraud Detection)

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

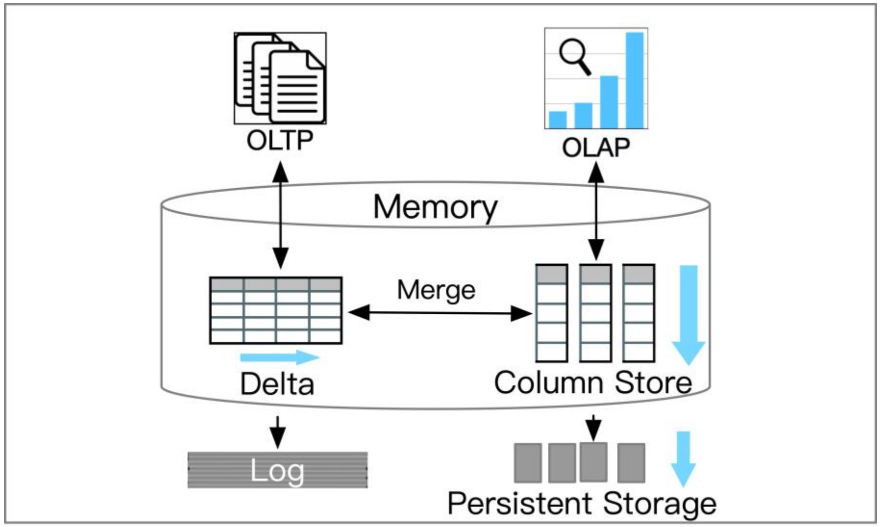
An Overview of HTAP Architectures

Case Study: SAP HANA



Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Challenges for HTAP Databases with architecture (d)



(d) Primary Column Store + Delta Row Store

Problems:

- Need to increase the TP scalability
- Need to increase workload isolation

Challenge: How to traverse the delta storage efficiently while keeping high throughput for HTAP

Possible solutions:

- Trade data freshness for AP throughput
- New Indexing techniques for delta traversal and delta merging

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

A summary of HTAP databases

Category	HTAP Databases	OLTP Throughput	OLAP Throughput	OLTP Scalability	OLAP Scalability	Workload Isolation	Data Freshness
Primary Row Store+ In Memory Column Store	Oracle Dual-Format SQL Server, DB2 BLU	High	High	Medium	Low	Low	High
Distributed Row Store + Column Store Replica	TiDB, F1 Lightning SingleStore	Medium	Medium	High	High	High	Low
Disk Row Store + Distributed Column Store	MySQL Heatwave, Oracle RAC	Medium	Medium	Medium	High	High	Medium
Primary Column Store + Delta Row Store	SAP HANA (without scale-out), Hyper	Medium	High	Low	Medium	Low	High

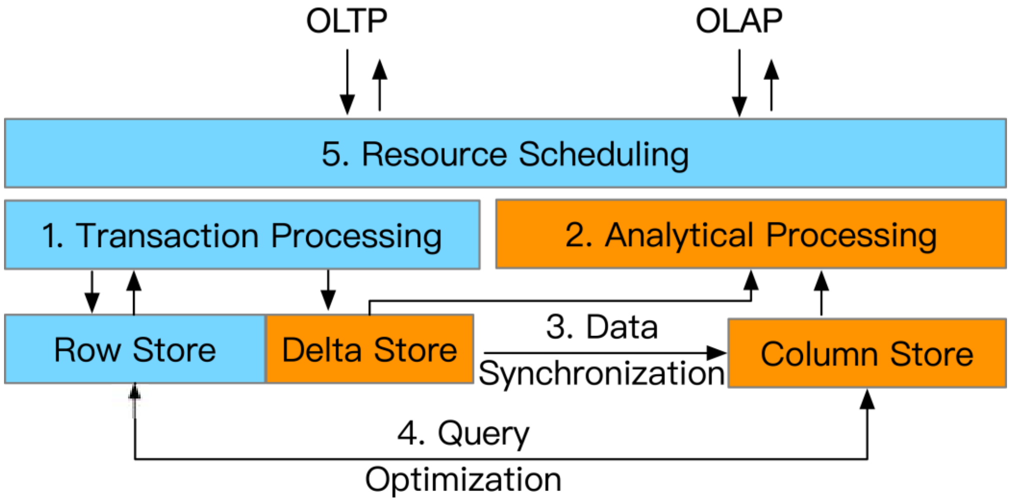
PURDUE UNIVERSITY®

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

HTAP TECHNIQUES

Overview of HTAP Techniques

1. **Transaction Processing:** updating the row store and writing the delta store
2. **Analytical Processing**： scanning the column store with delta store
3. **Data Synchronization:** merging the delta data to column store
4. **Query Optimization:** planning queries against row store and column store
5. **Resource Scheduling:** scheduling resources for OLTP and OLAP instances



Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Transaction Processing

1. **Standalone Transaction Processing with In-Memory Delta Update**
    - Standalone transaction processing with MVCC protocol
    - In-memory delta update for insert/delete/update operations
    - E.g., Oracle, SQL Server, SAP HANA

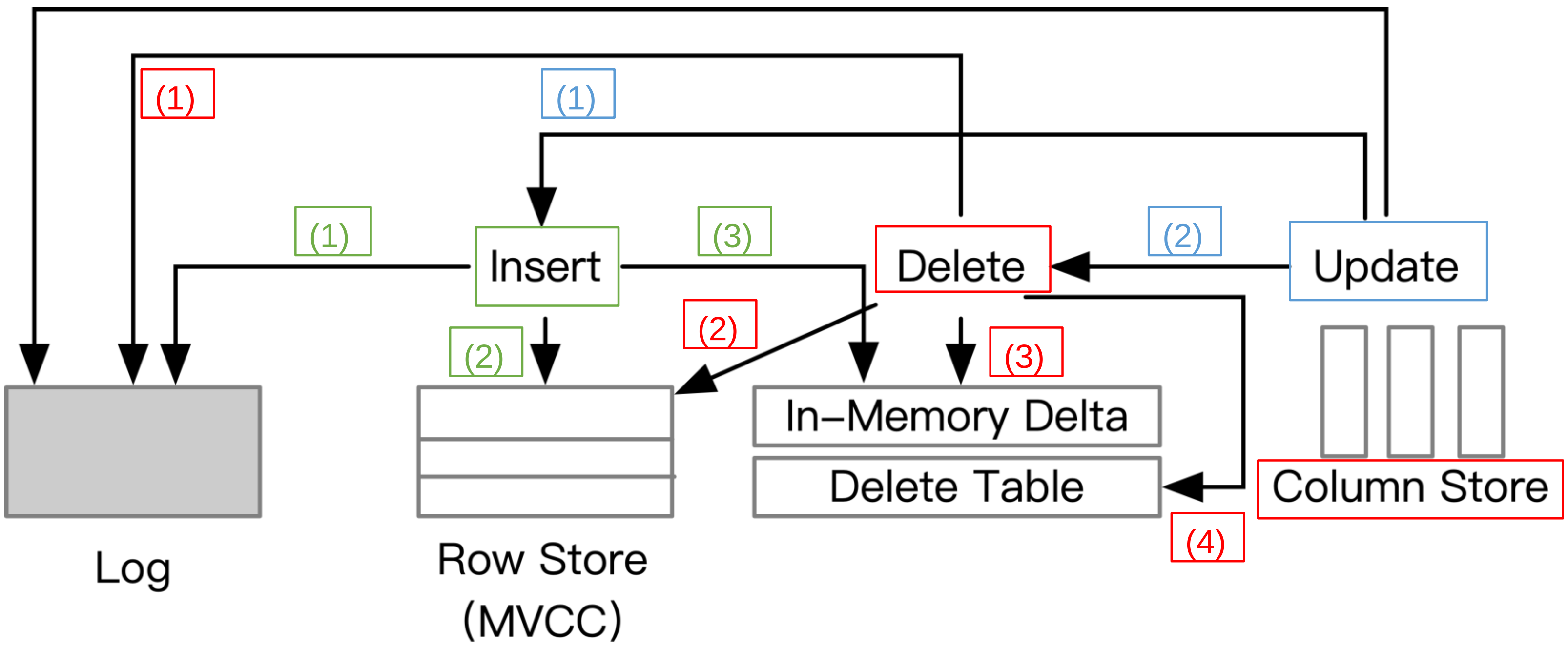
2. **Distributed Transaction Processing with Log Replay**
    - Raft protocol for distributed TP and data replication
    - Log replay for updating the row store and column store
    - E.g., F1 Lightning, TiDB

    Master-slave replication for distributed TP, e.g., Singlestore

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Transaction Processing

1. **Standalone TP for insert/delete/update operations**



Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Transaction Processing

Three implementations for an in-memory delta store:

1. Heap table
2. Index organized table
3. L1 cache

Delta Store	Databases	Pros	Cons
Heap table	Oracle	Fast Insertion	Slow Lookup
Index Organized Table	SQL Server	Fast Lookup	Slow Insertion
L1 Cache	SAP HANA	Fast Insertion	Slow Lookup Low Capacity

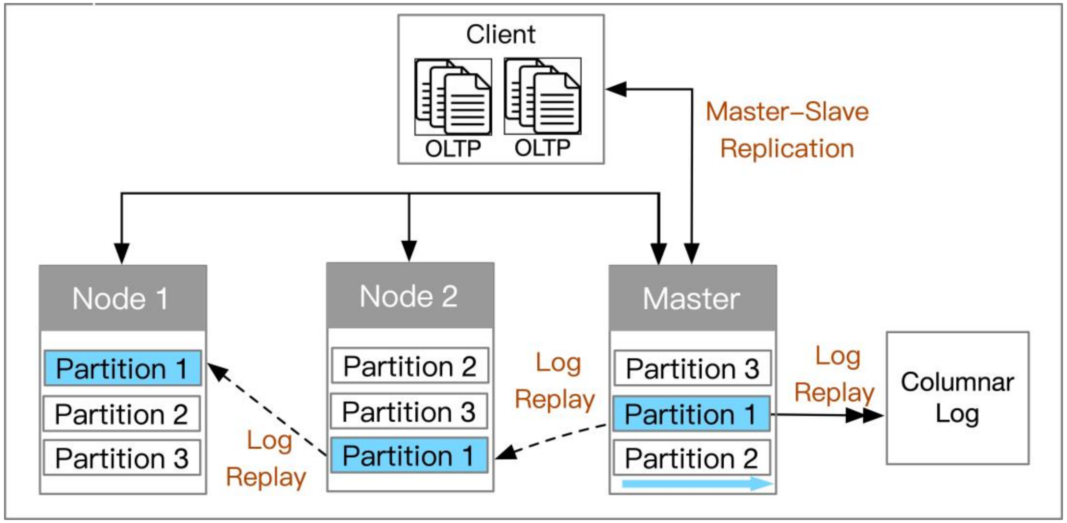
Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Transaction Processing

Distributed Transaction Processing with Log Replay (1)

Master-slave replication

- Master node handles the transactions, then replicate the logs to slave nodes
- E.g., SingleStore



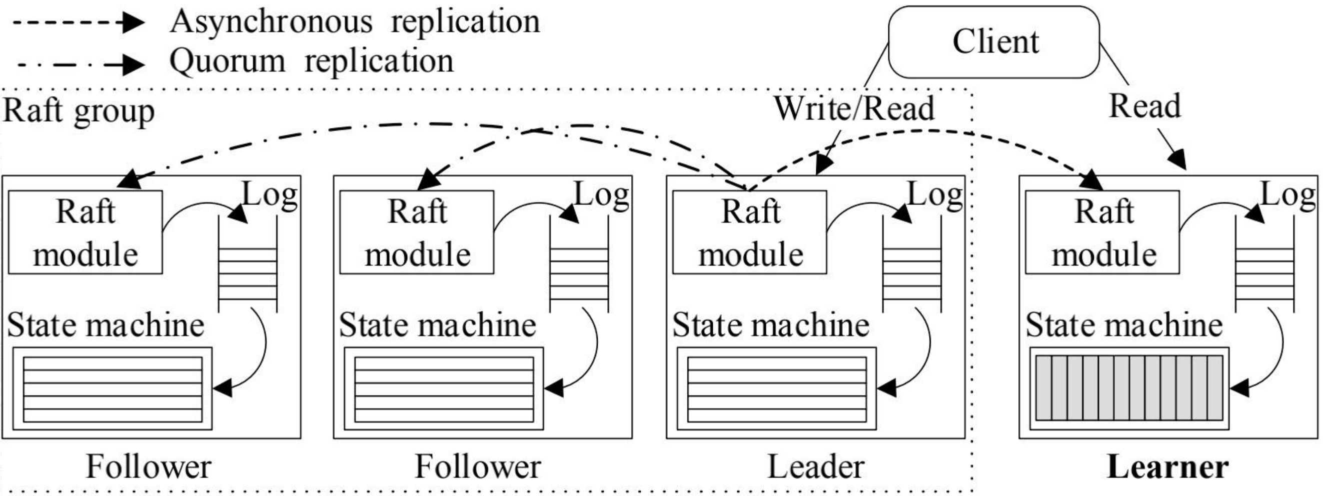
Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Transaction Processing

Distributed Transaction Processing with Log Replay (2)

Modified Raft Protocol for TP and AP nodes

- Leader (row), Follower (row), Learner (column)



Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Comparisons of TP techniques in HTAP Databases

Transaction Processing Type	Databases	TP Techniques	Delta	Pros	Cons
Standalone TP + In-memory Delta Update	Oracle, SQL Server	MVCC	In-Memory Delta	High Efficiency	Low Scalability
Distributed TP + Log Replay	SingleStore	Master-Slave Replication	Log Files	High Efficiency	Low Freshness
	TiDB, F1 Lightning	2PC+Paxos	Log Files	High Scalability	Low Efficiency

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Analytical Processing

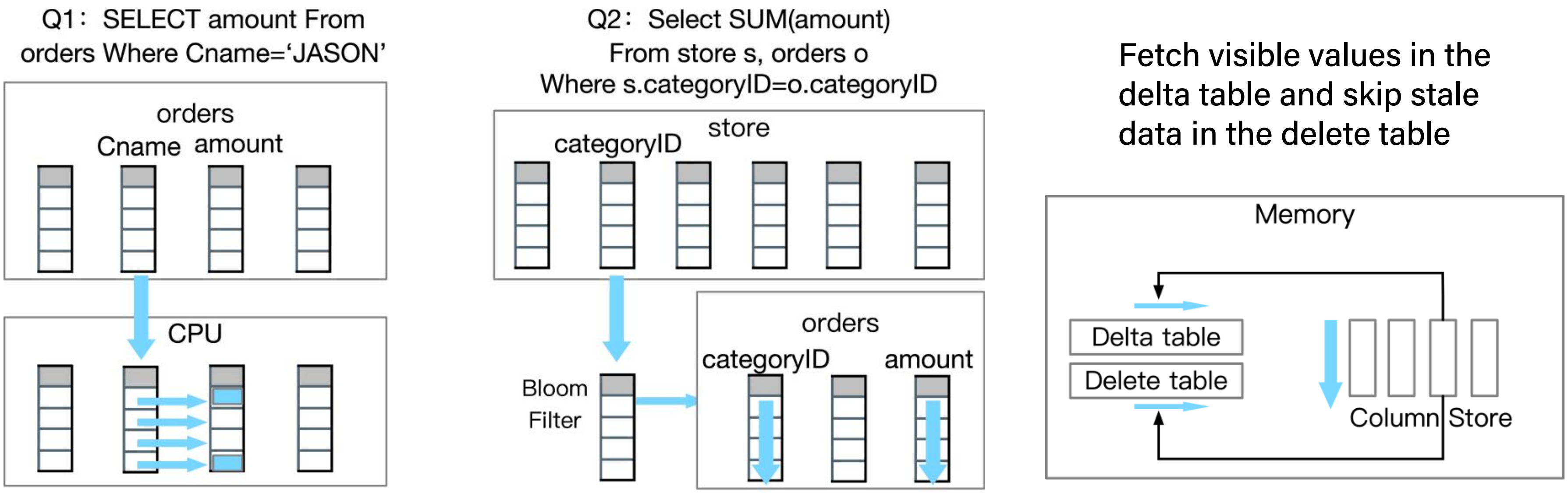
1. **Standalone Columnar Scan with In-Memory Delta Traversing**
   - Single Instructions Multiple Data (SIMD), Vector Processing
   - In-Memory Delta Traversing
   - E.g., Oracle, SQL Server

2. **Distributed Columnar Scan with Log File Scanning**
   - Distributed Query Processing over Columnar Segments
   - Disk-based Log Files Merging and Scanning
   - E.g., F1 Lightning, TiDB

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Analytical Processing

1. Standalone Columnar Scan with In-Memory Delta Traversing


Q1: SELECT amount From orders Where Cname='JASON'
orders
Cname
amount
CPU
Q2: Select SUM(amount) From store s, orders o Where s.categoryID=o.categoryID
store
categoryID
Bloom Filter
orders
categoryID
amount
Memory
Delta table
Delete table
Column Store

Fetch visible values in the delta table and skip stale data in the delete table

(a) SIMD query processing

(b) Vector join based on a bloom filter

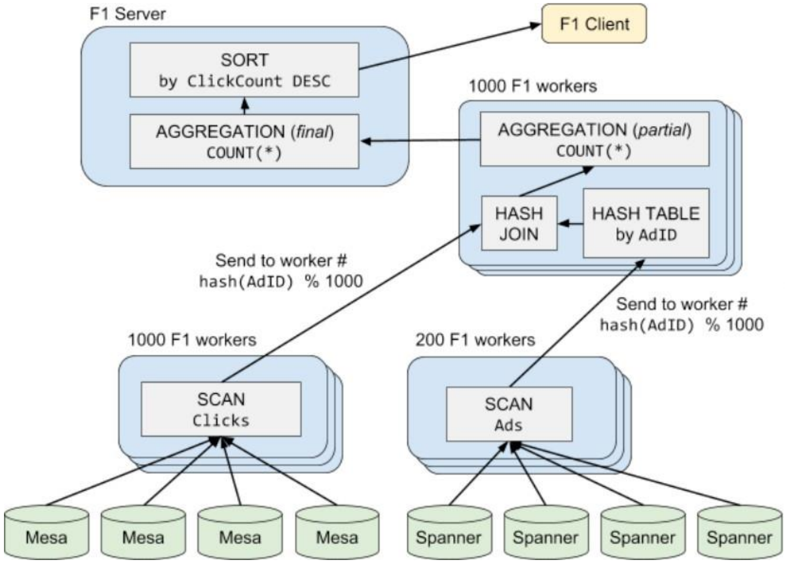
(c) Column Scan with delta traversing

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

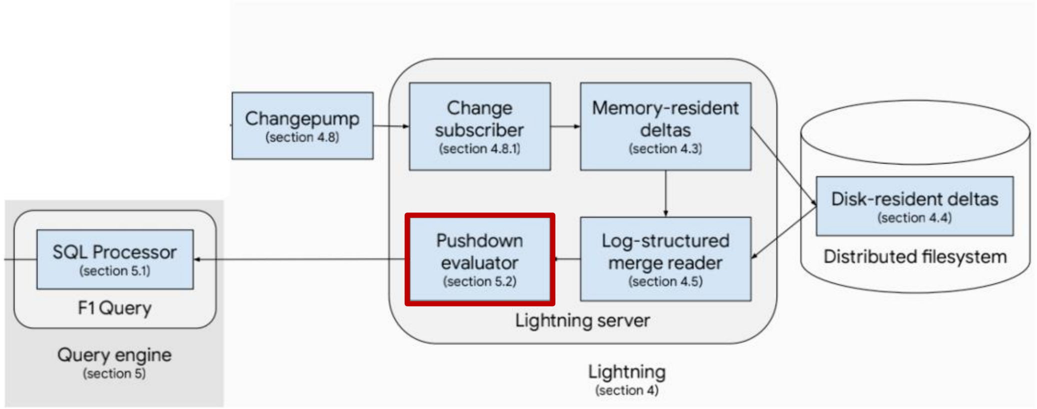
Analytical Processing

2. Distributed Columnar Scan with Log File Scanning

Distributed Columnar Scan



Log File Scanning



Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Comparisons of AP techniques in HTAP Databases

Analytical Processing Type	Databases	AP techniques	Delta	Pros	Cons
Standalone Columnar Scan + In-Memory Delta Traversing	Oracle, SQL Server, SAP HANA	Vector query processing + Delta traversing	In-memory delta table	High Freshness	Large Memory Size
Distributed Columnar Scan + Log File Scanning	TiDB, F1 Lightning	Distributed query processing + Log scanning	Disk-based log files	High Scalability	Low Efficiency

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Data Synchronization

Periodically merge the latest transaction data to the column store

- **Type 1: In-memory delta merge**
    1. Threshold-based merging
    2. Two-phase data migration
    3. Dictionary-based migration

  Oracle, SQL Server, SAP HANA

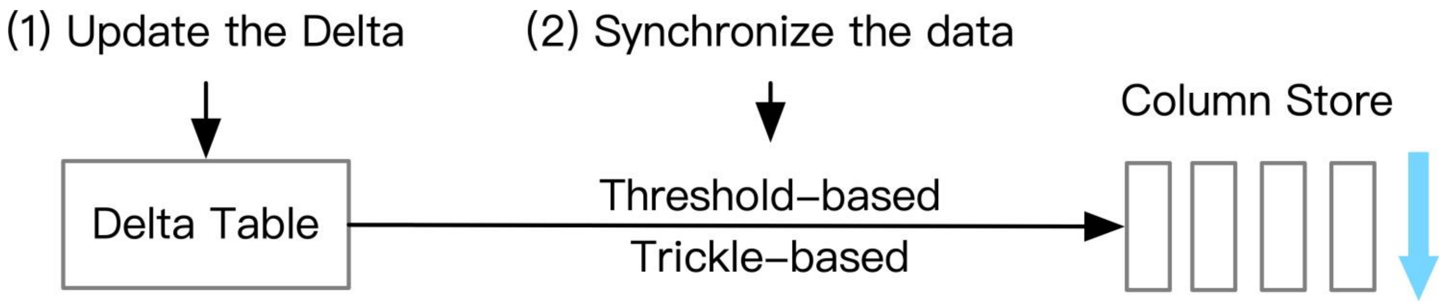
- **Type 2: Log-based delta merge**
    1. LSM-tree and B-tree

  TiDB, F1 Lightning

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Data Synchronization

1. **In-Memory Delta Merging**
   - **Method 1: Threshold-based merging**
   - e.g., threshold reaches 90% of column store

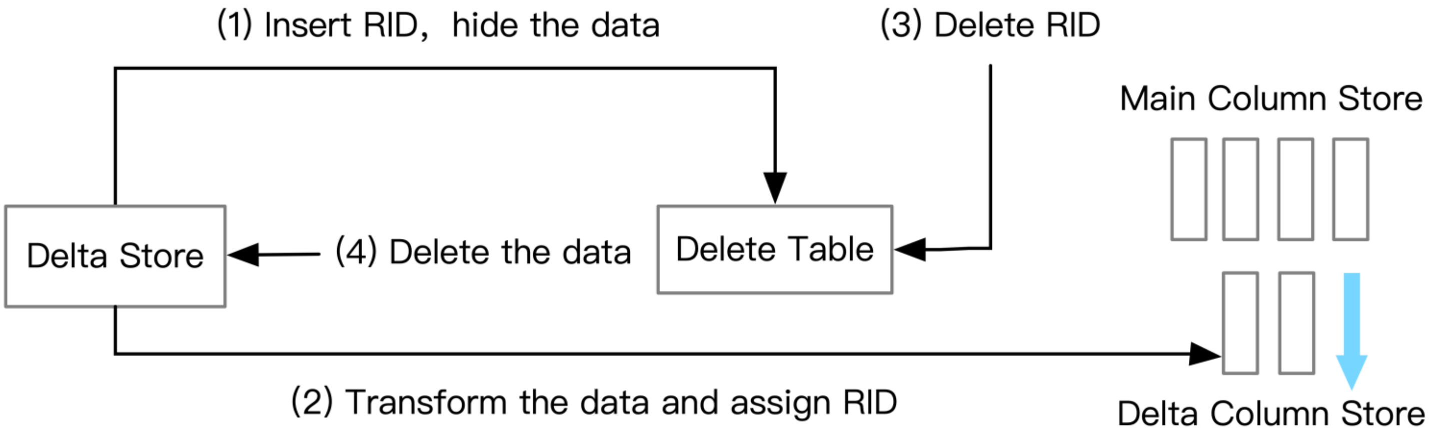


(a) Threshold-based merging

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Data Synchronization

1. **In-Memory Delta Merging**
   - **Method 2: Two-phase delta migration**
   - Phase 1: Preparation on migration
   - Phase 2: Operation on migration

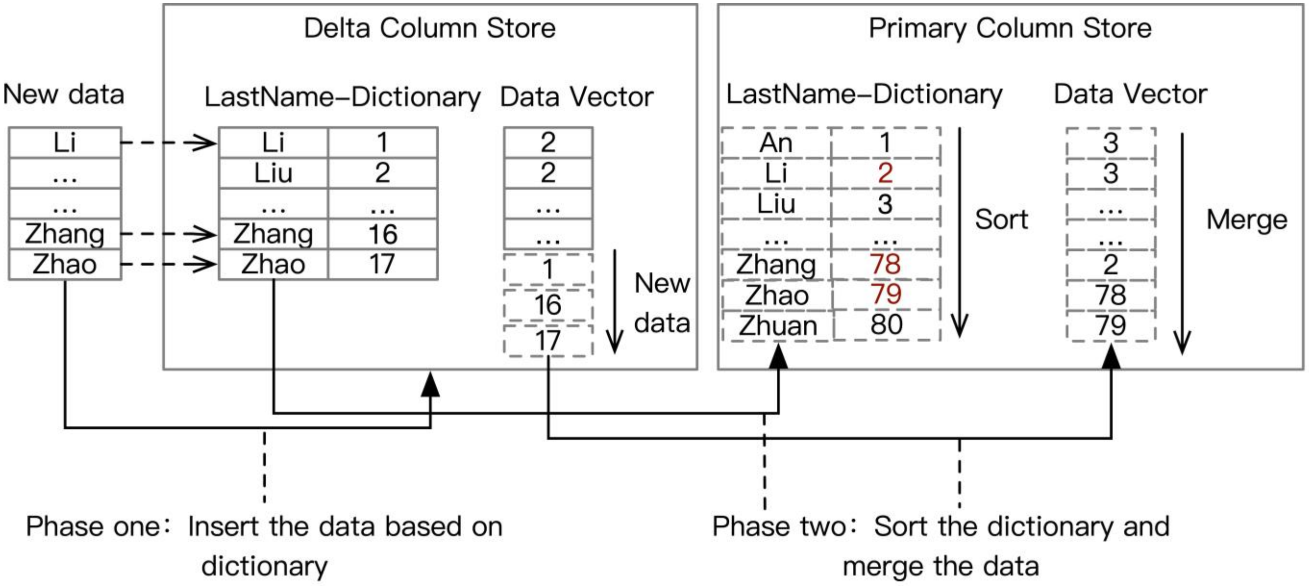


(b) Two-phase data migration

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Data Synchronization

1. **In-Memory Delta Merging**
   - **Method 3: Dictionary-based merging**



Delta Column Store

New data

Li
...
...
Zhang
Zhao

LastName–Dictionary

Li	1
Liu	2
...	...
Zhang	16
Zhao	17

Data Vector

2
2
...
...
1
16
17

New data

Primary Column Store

LastName–Dictionary

An	1
Li	2
Liu	3
...	...
Zhang	78
Zhao	79
Zhuan	80

Sort

Data Vector

3
3
...
...
2
78
79

Merge

Phase one：Insert the data based on dictionary

Phase two：Sort the dictionary and merge the data

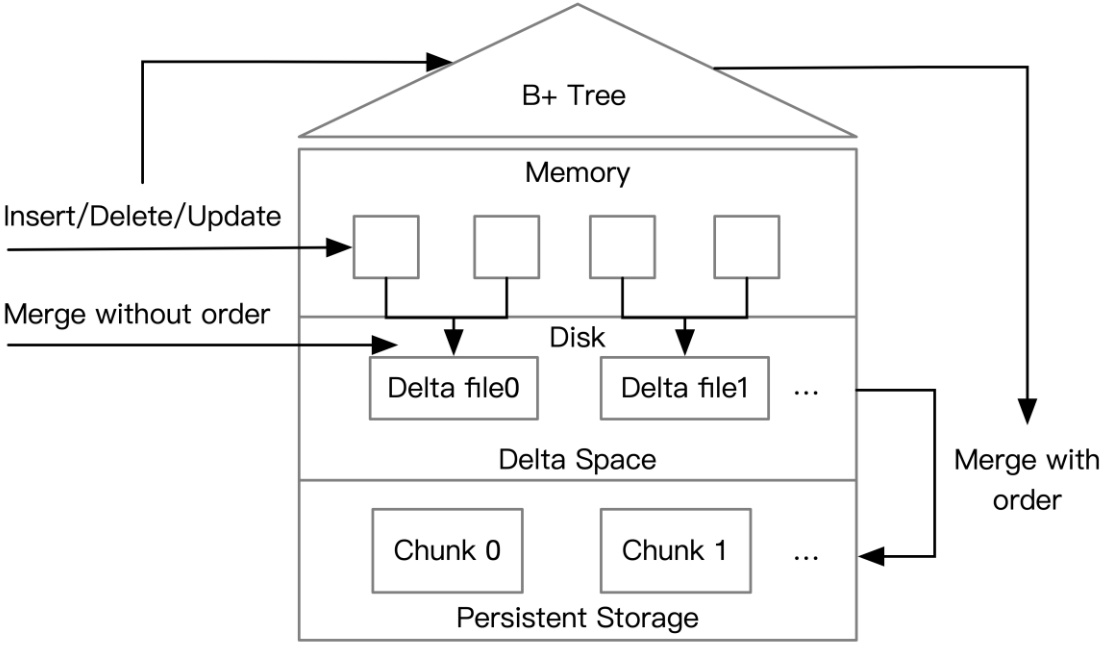
(c) Dictionary-based merging

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Data Synchronization

2. Log-based delta merge

- Memory-resident deltas (row-wise)
- Disk-resident deltas (column-wise)



- Merging and Collapsing
- B+-Tree for fast merging

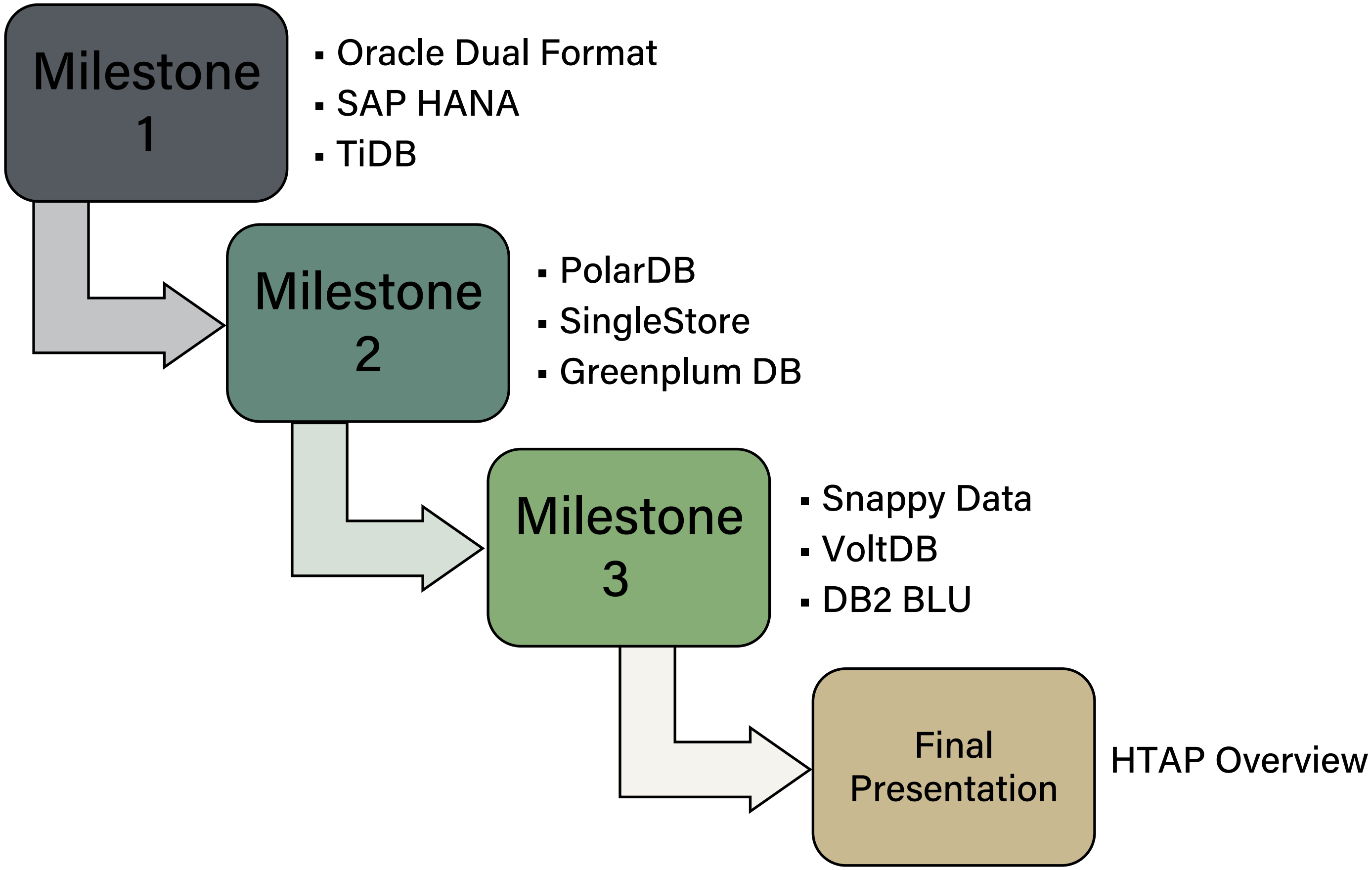
Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Comparisons of DS techniques in HTAP Databases

Data Synchronization	Databases	DS techniques	Pros	Cons
In-Memory delta merge	Oracle, SQL Server, SAP HANA	• Threshold-based merging • Two-phase delta migration • Dictionary-based merging	High Efficiency	Low Scalability
Log-based delta merge	TiDB, F1 Lightning	• Multi-level deltas • B+tree • Log merging	High Scalability	High Merge Cost

Source: HTAP Databases: What is New and What is Next, SIGMOD '22: Proceedings of the 2022 International Conference on Management of Data

Project Summary



Milestone 1

- Oracle Dual Format
- SAP HANA
- TiDB

Milestone 2

- PolarDB
- SingleStore
- Greenplum DB

Milestone 3

- Snappy Data
- VoltDB
- DB2 BLU

Final Presentation

HTAP Overview

THANK YOU

PURDUE
UNIVERSITY®