PingCAP

Big Data in TiDB

shenli@PingCAP

PingCAP

About me

- Shen Li (申砾)
- Tech Lead of TiDB, VP of Engineering
- Netease / 360 / PingCAP
- Infrastructure software engineer

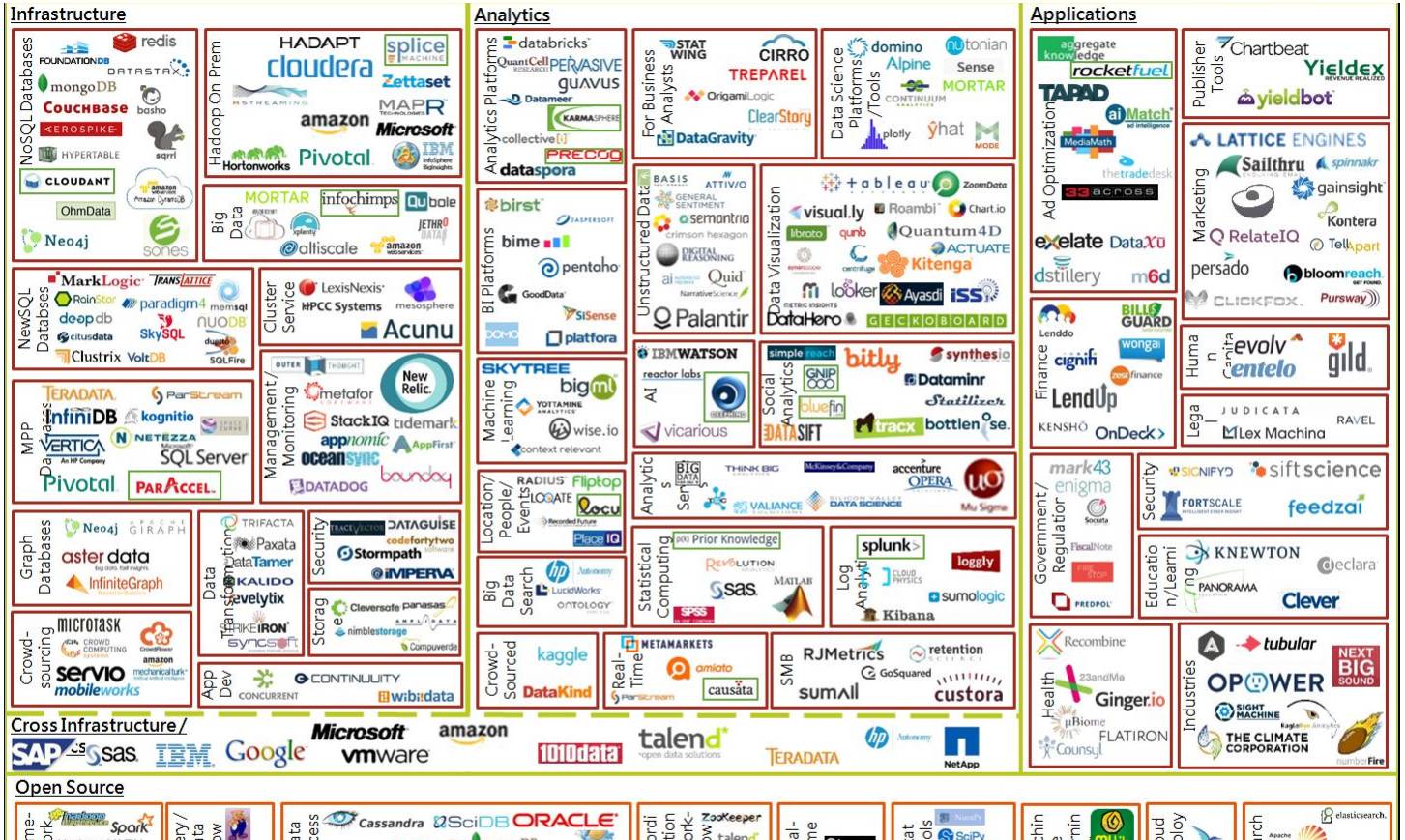
What is Big Data?

Big Data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.

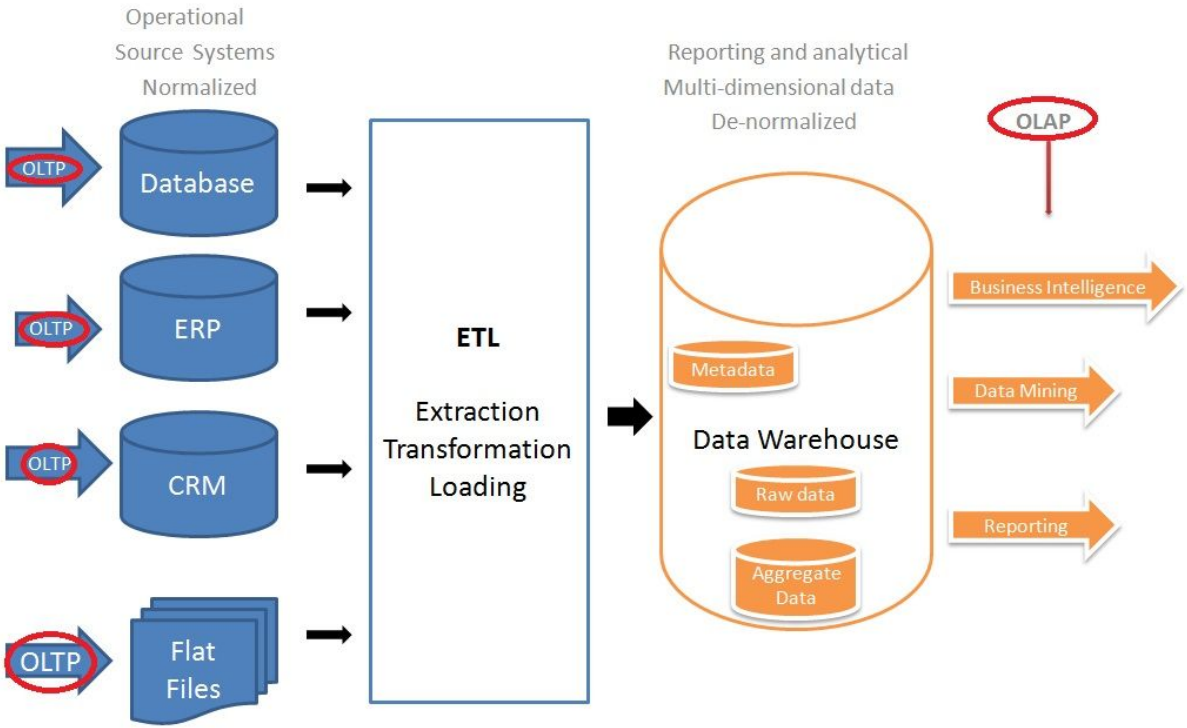
---- From Wikipedia

PingCAP

Big Data Landscape



OLTP and OLAP



OLTP (Online Transaction Processing) versus
OLAP (Online Analytical Processing)

What is TiDB

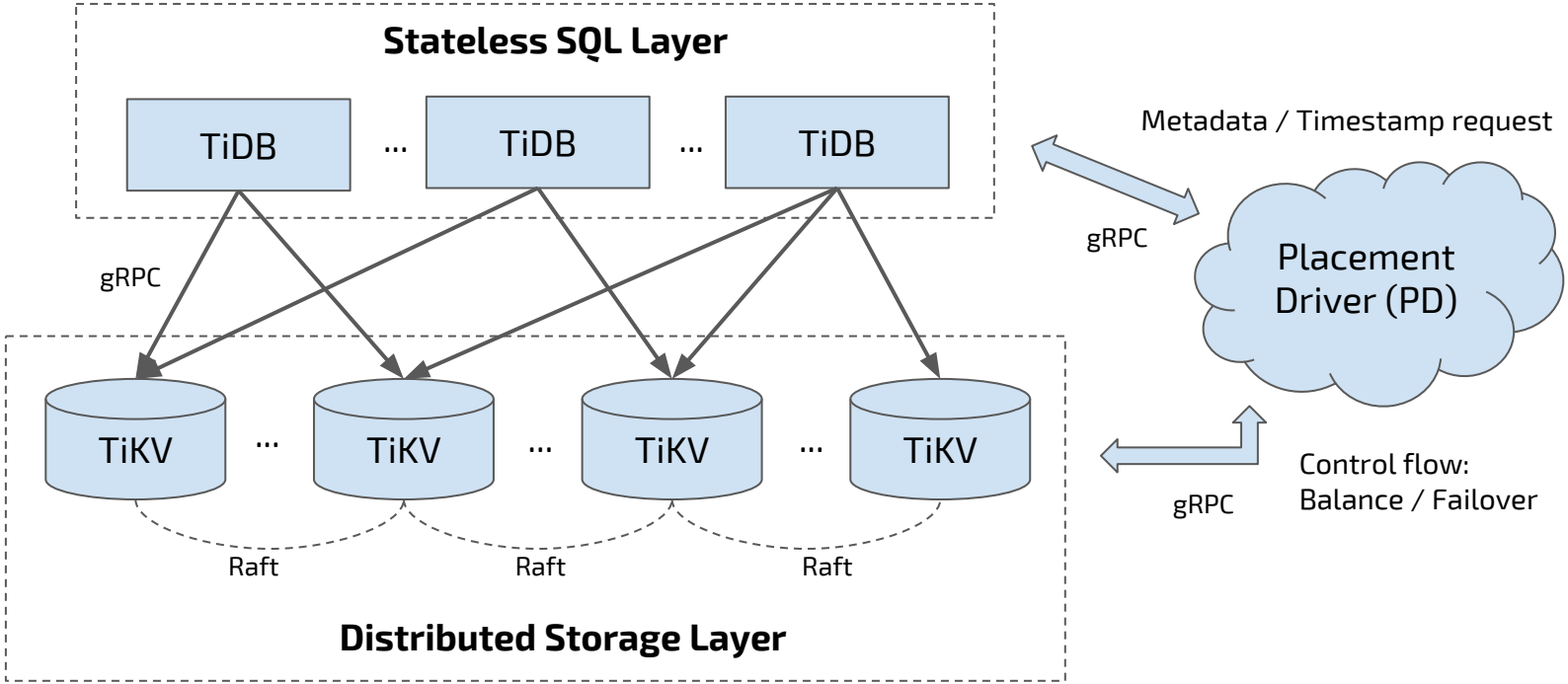
- SQL is necessary
- Scale is easy
- Compatible with MySQL, at most cases
- OLTP + OLAP = HTAP (Hybrid Transactional/Analytical Processing)
  - Transaction + Complex query
- 24/7 availability, even in case of datacenter outages
  - Thanks to Raft consensus algorithm
- Open source, of course.

TiDB

A Distributed SQL Database

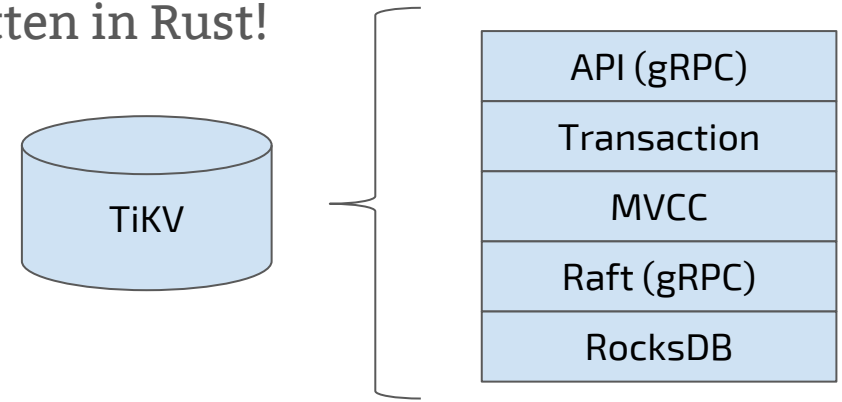
PingCAP

Architecture



Storage stack 1/2

- TiKV is the underlying storage layer
- Physically, data is stored in RocksDB
- We build a Raft layer on top of RocksDB
  - What is Raft?
- Written in Rust!

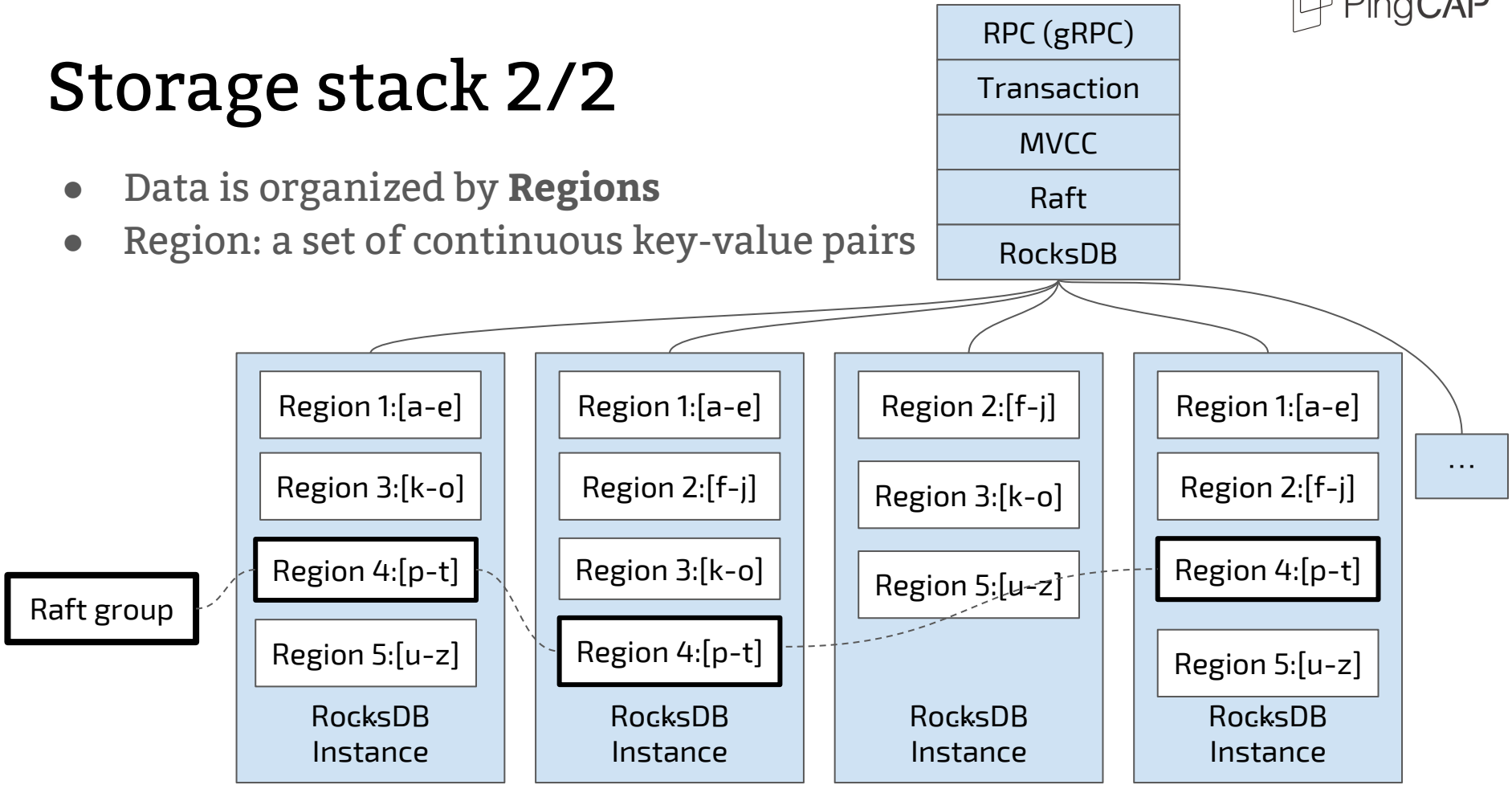


Transactional KV API
(https://github.com/pingcap/tidb/blob/master/cmd/benchkv/main.go)

Raw KV API
(https://github.com/pingcap/tidb/blob/master/cmd/benchraw/main.go)

Storage stack 2/2

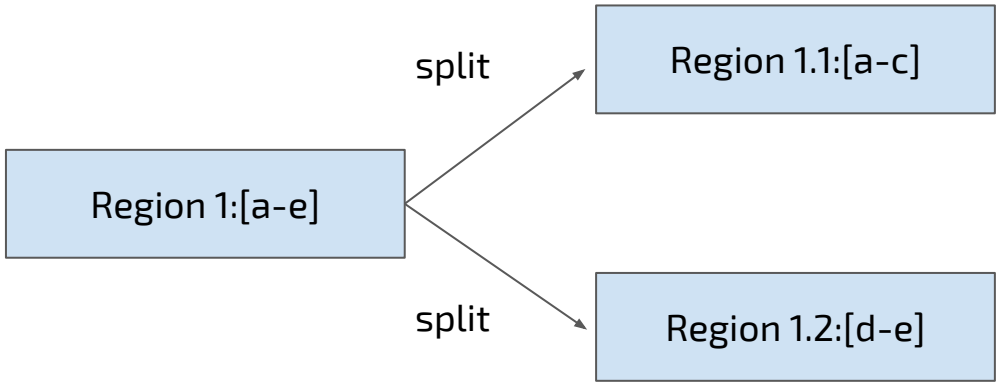
- Data is organized by **Regions**
- Region: a set of continuous key-value pairs



PingCAP

Dynamic Multi-Raft

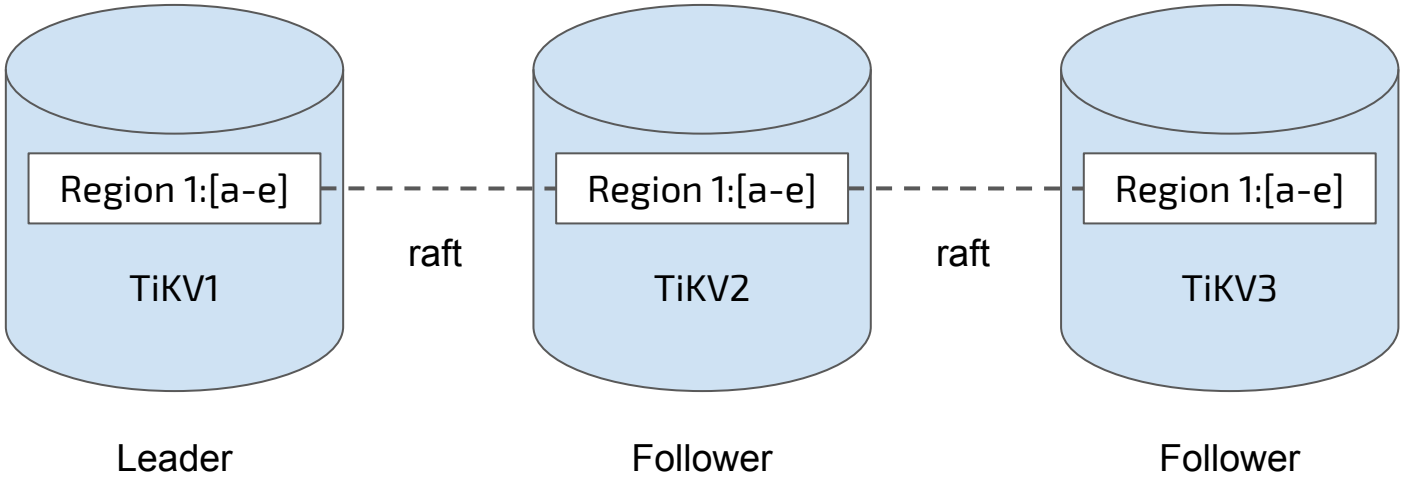
- What's Dynamic Multi-Raft?
  - Dynamic split / merge
- Safe split / merge



PingCAP

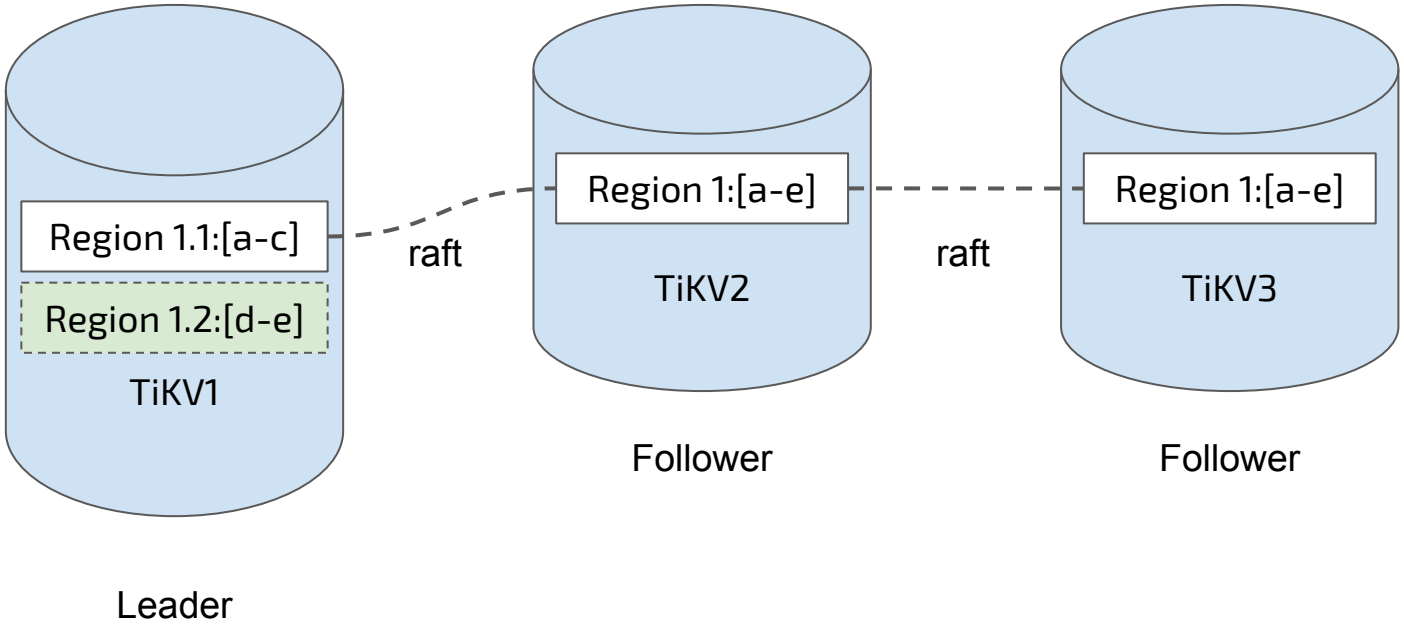
Safe Split: 1/4

Raft group

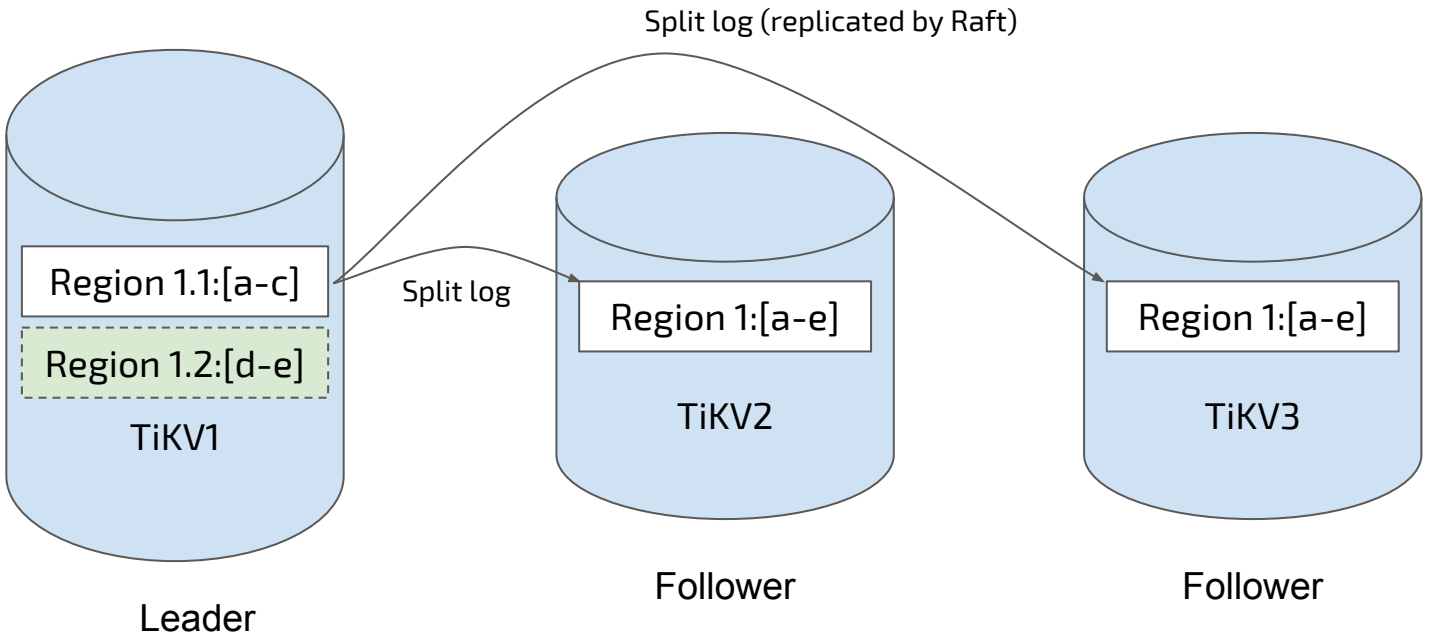


PingCAP

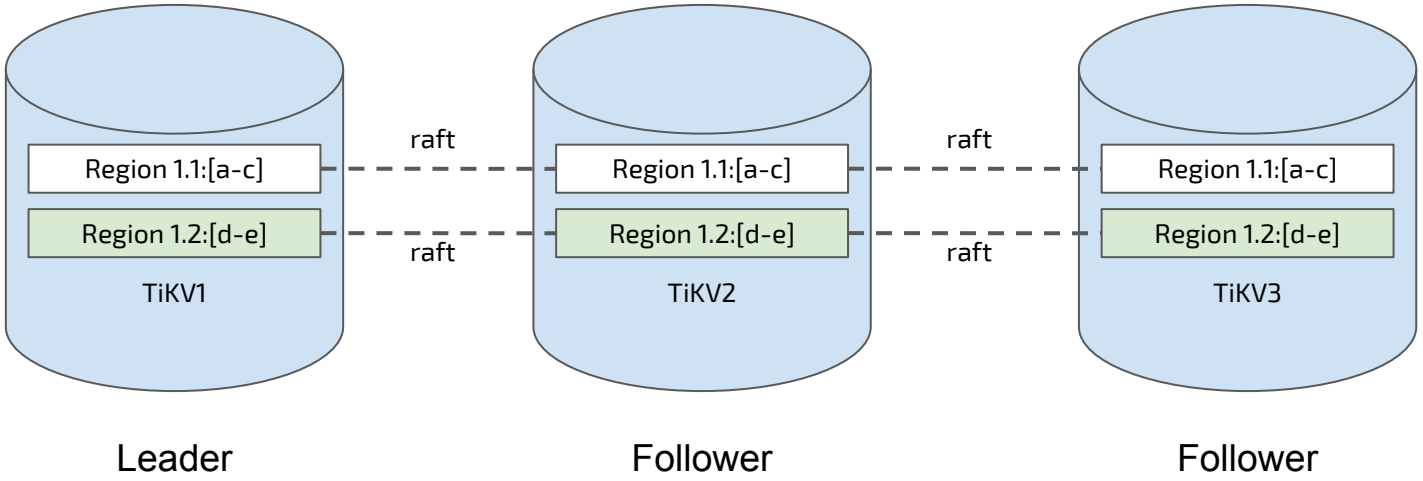
Safe Split: 2/4



Safe Split: 3/4

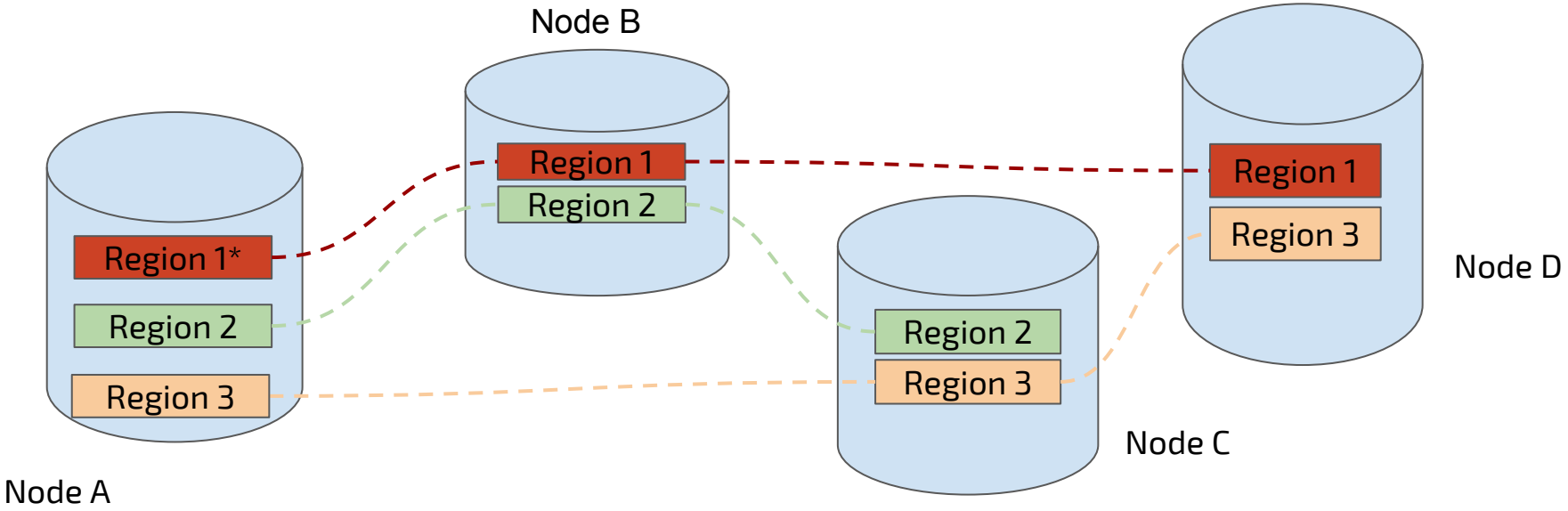


Safe Split: 4/4



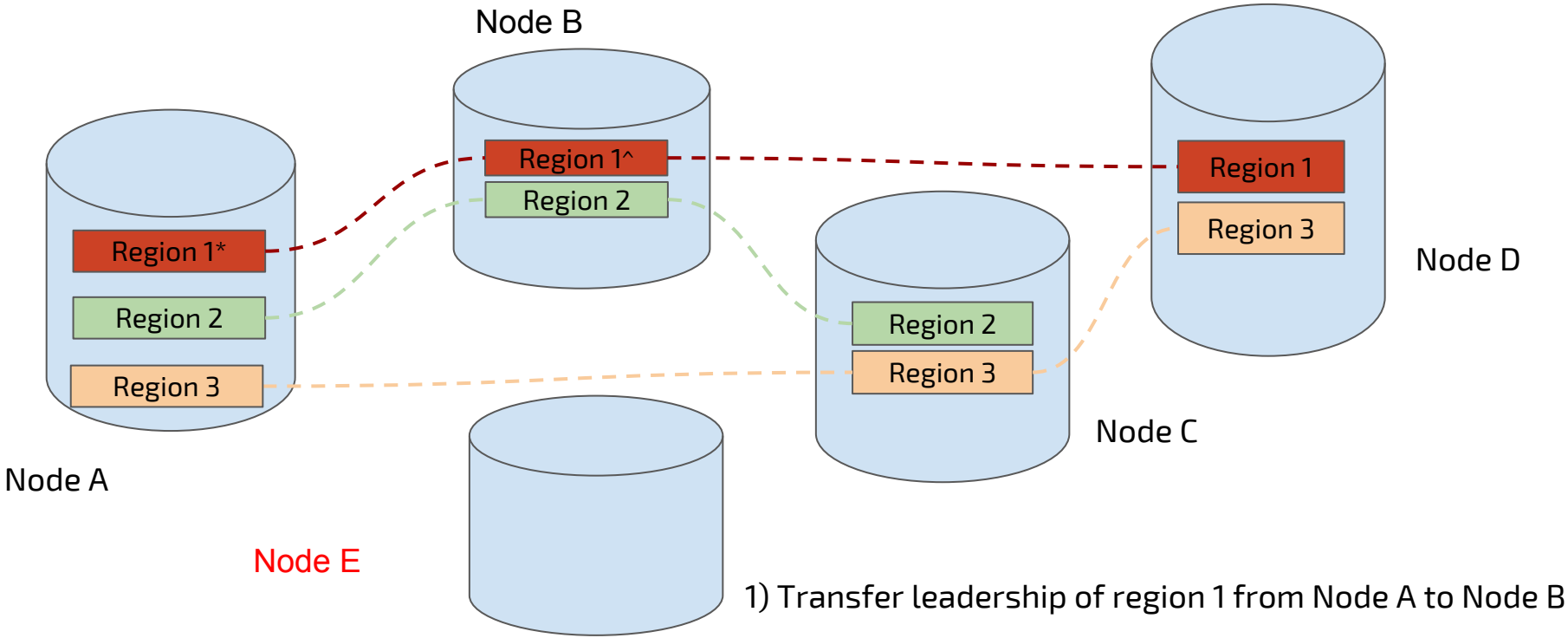
PingCAP

Scale-out (initial state)



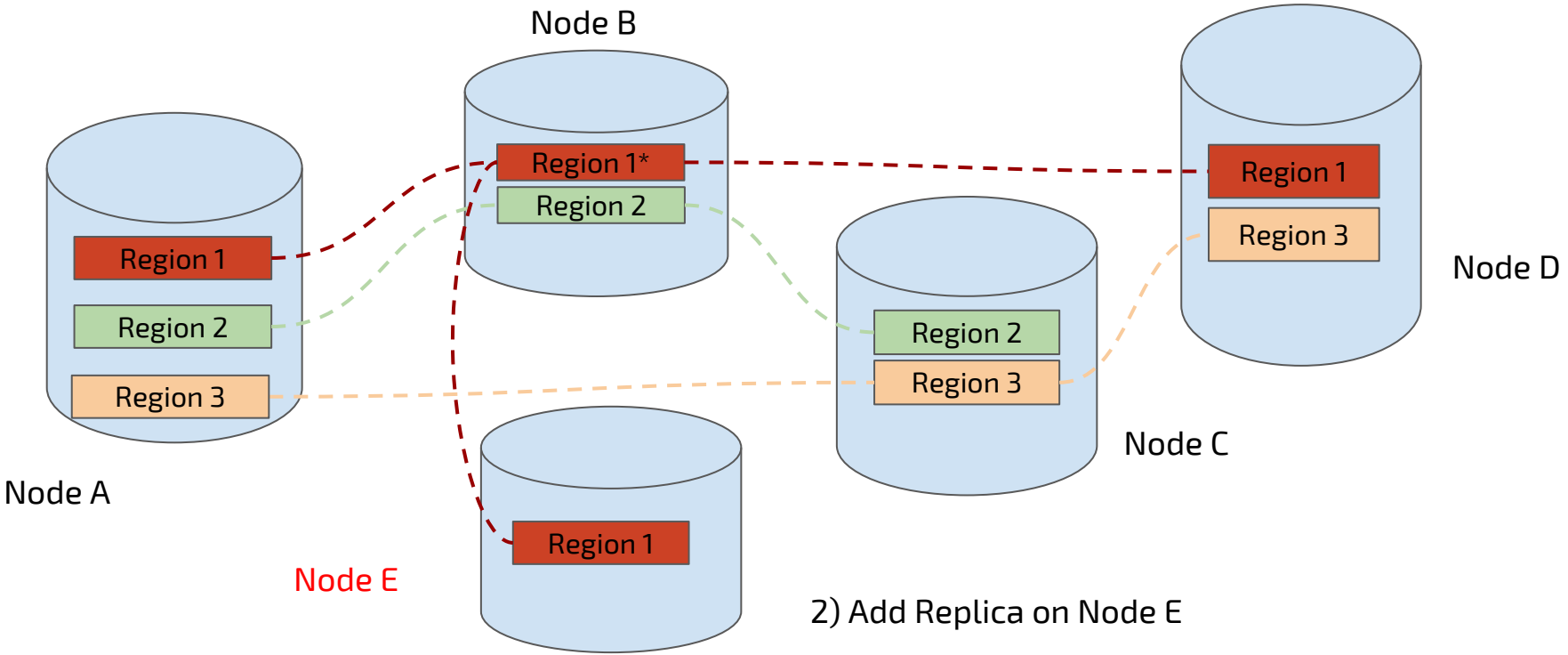
PingCAP

Scale-out (add new node)



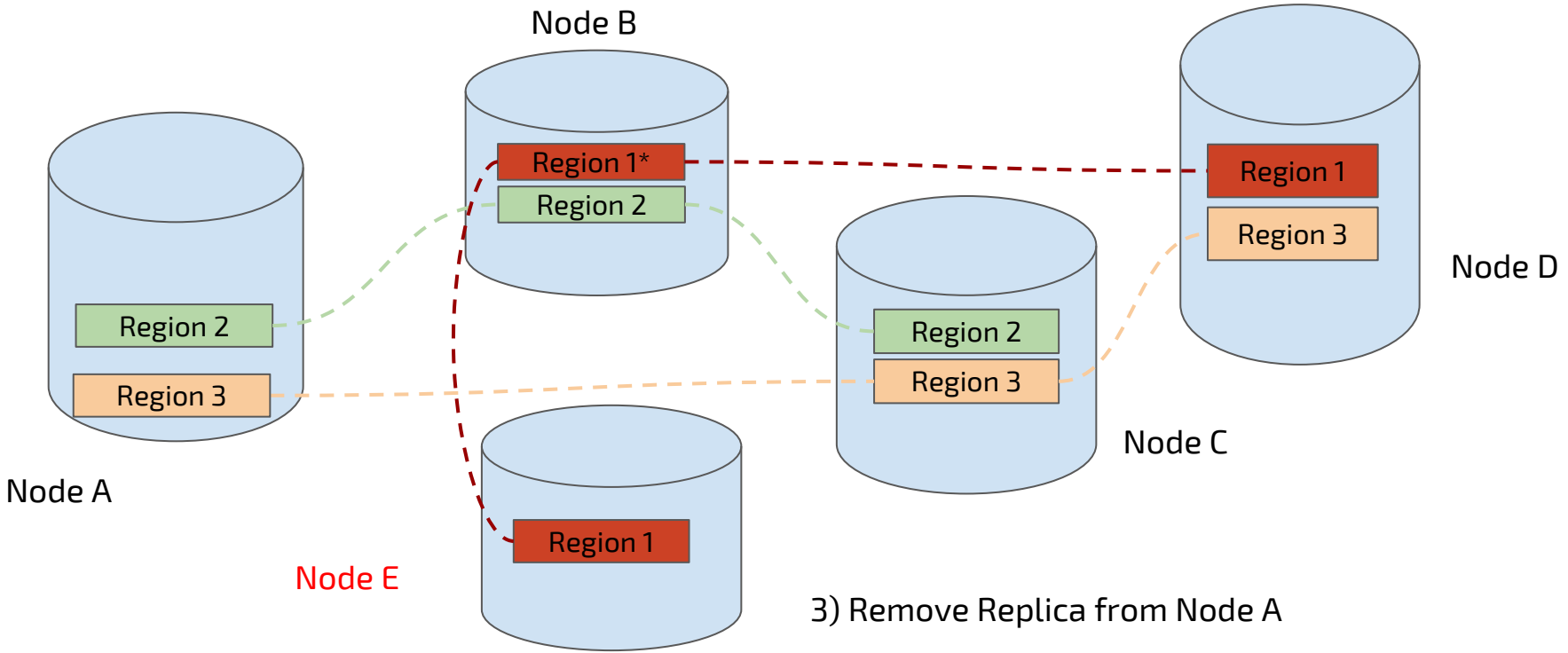
PingCAP

Scale-out (balancing)



PingCAP

Scale-out (balancing)



PingCAP

ACID Transaction

- Based on Google Percolator
- ‘Almost’ decentralized 2-phase commit
  - Timestamp Allocator
- Optimistic transaction model
- Default isolation level: Repeatable Read
- External consistency: Snapshot Isolation + Lock
    - SELECT ... FOR UPDATE

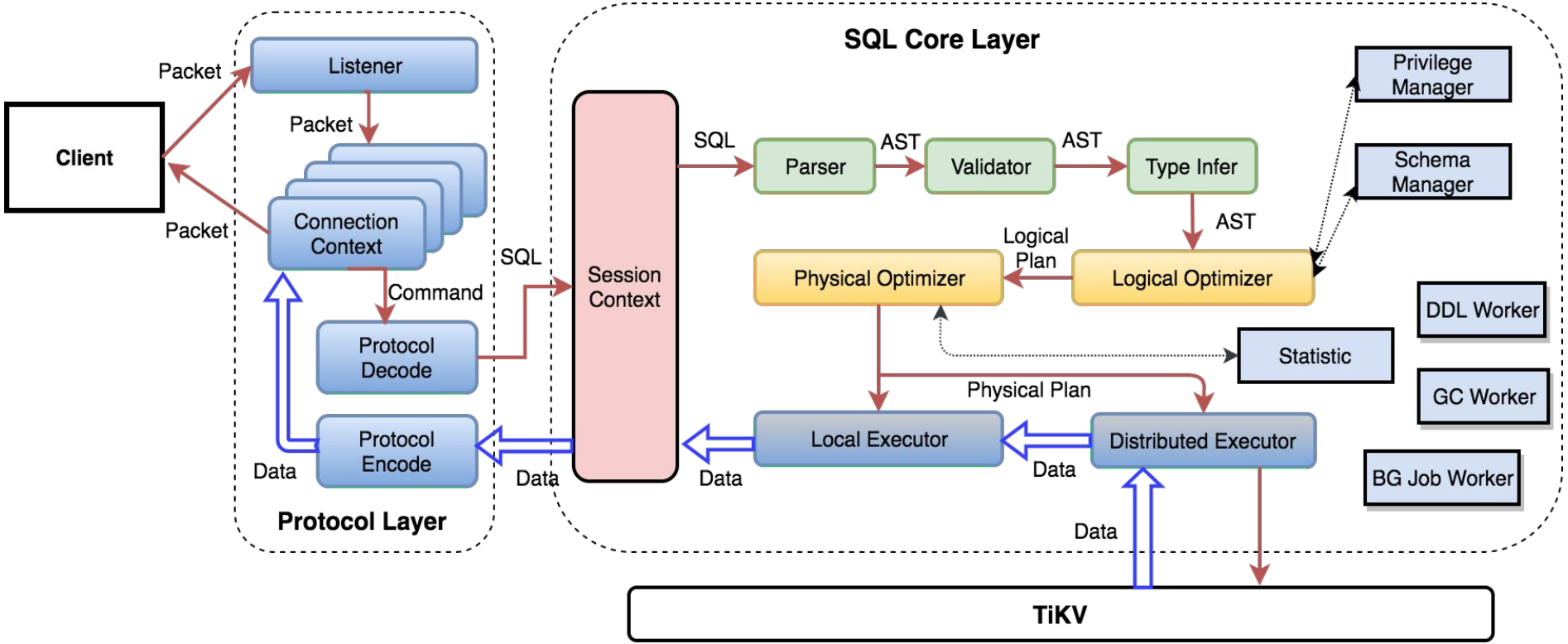
PingCAP

Distributed SQL

- Full-featured SQL layer
- Predicate pushdown
- Distributed join
- Distributed cost-based optimizer (Distributed CBO)

PingCAP

TiDB SQL Layer overview



PingCAP

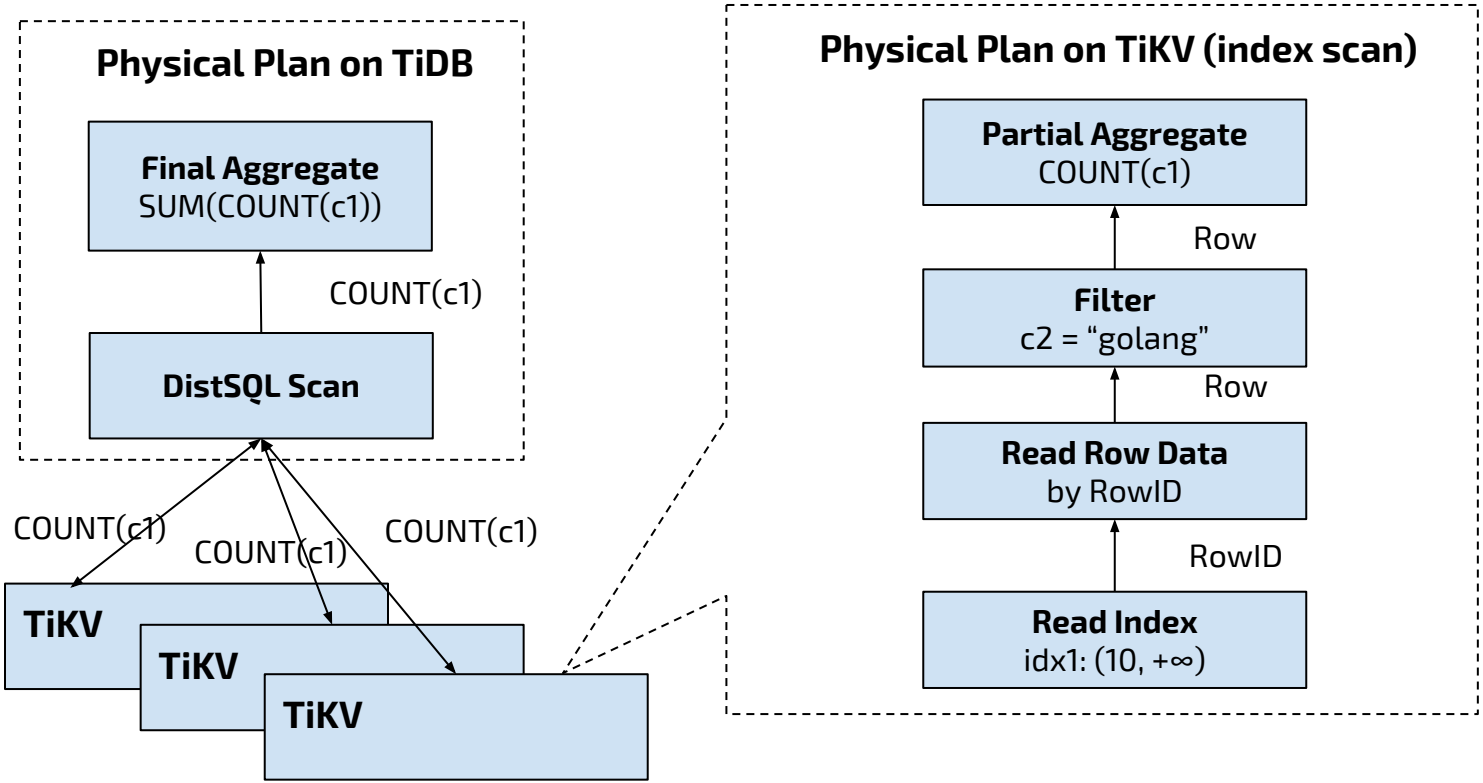
What happens behind a query

```
CREATE TABLE t (c1 INT, c2 TEXT, KEY idx_c1(c1));

SELECT COUNT(c1) FROM t WHERE c1 > 10 AND c2 = ‘golang’;
```

PingCAP

Query Plan



PingCAP

What happens behind a query

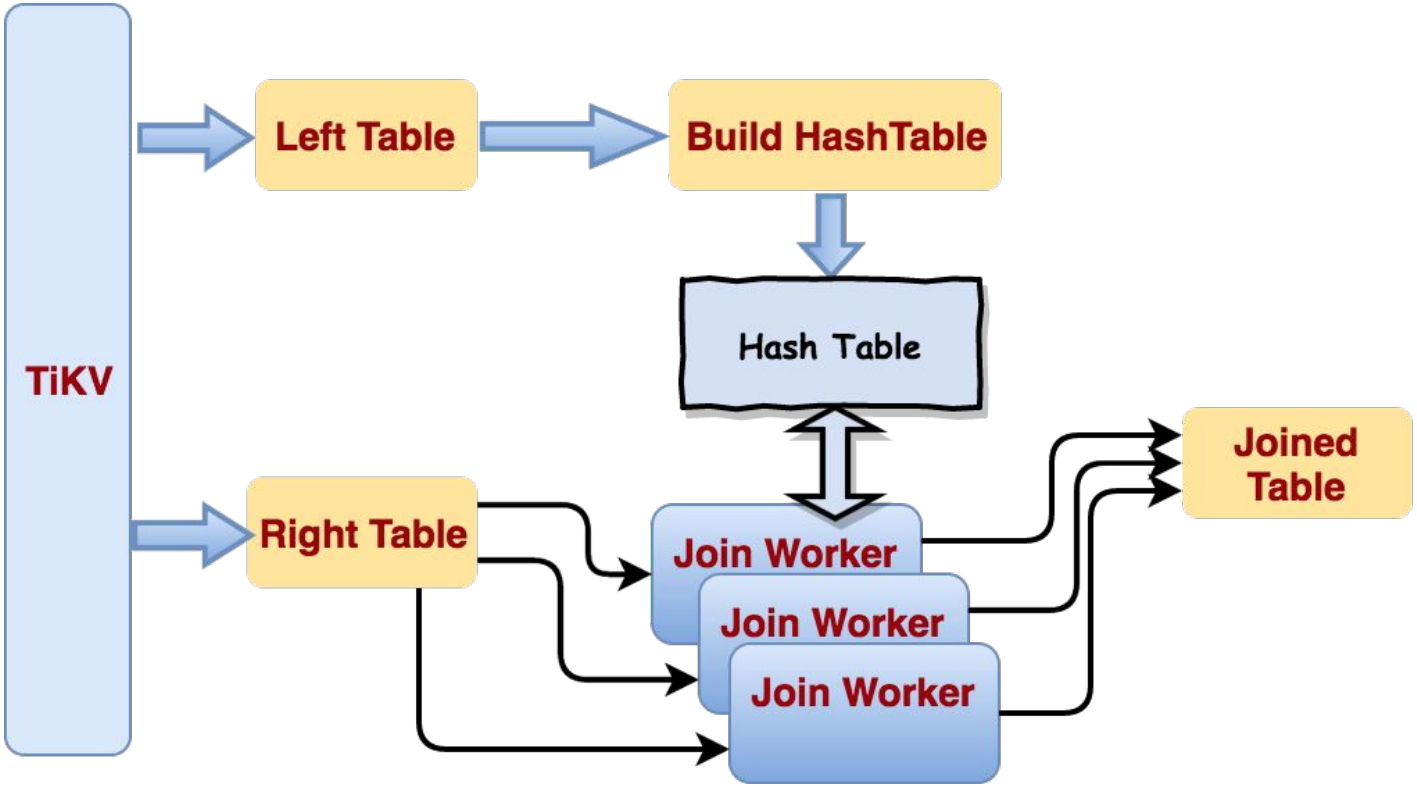
```
CREATE TABLE left (id INT, email TEXT, KEY idx_id(id));

CREATE TABLE right (id INT, email TEXT, KEY idx_id(id));

SELECT * FROM left join right WHERE left.id = right.id;
```

PingCAP

Distributed Join (HashJoin)



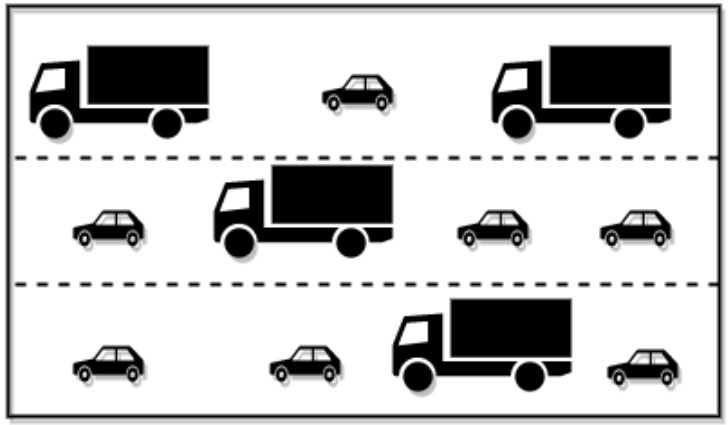
PingCAP

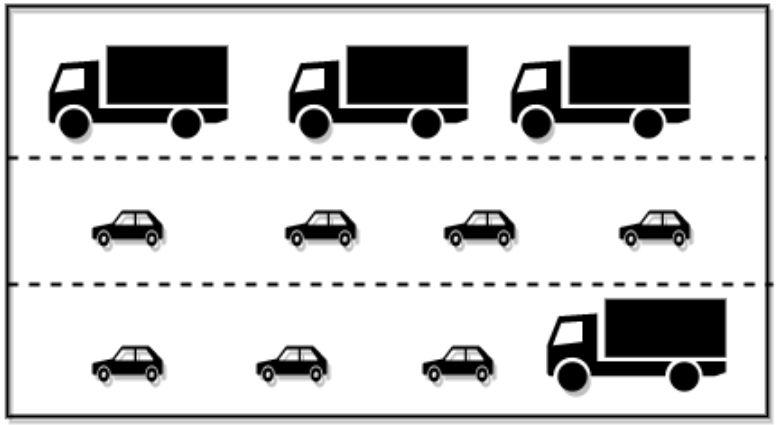
Supported Distributed Join Type

- Hash Join
- Sort merge Join
- Index-lookup Join

PingCAP

Hybrid Transactional/Analytical Processing





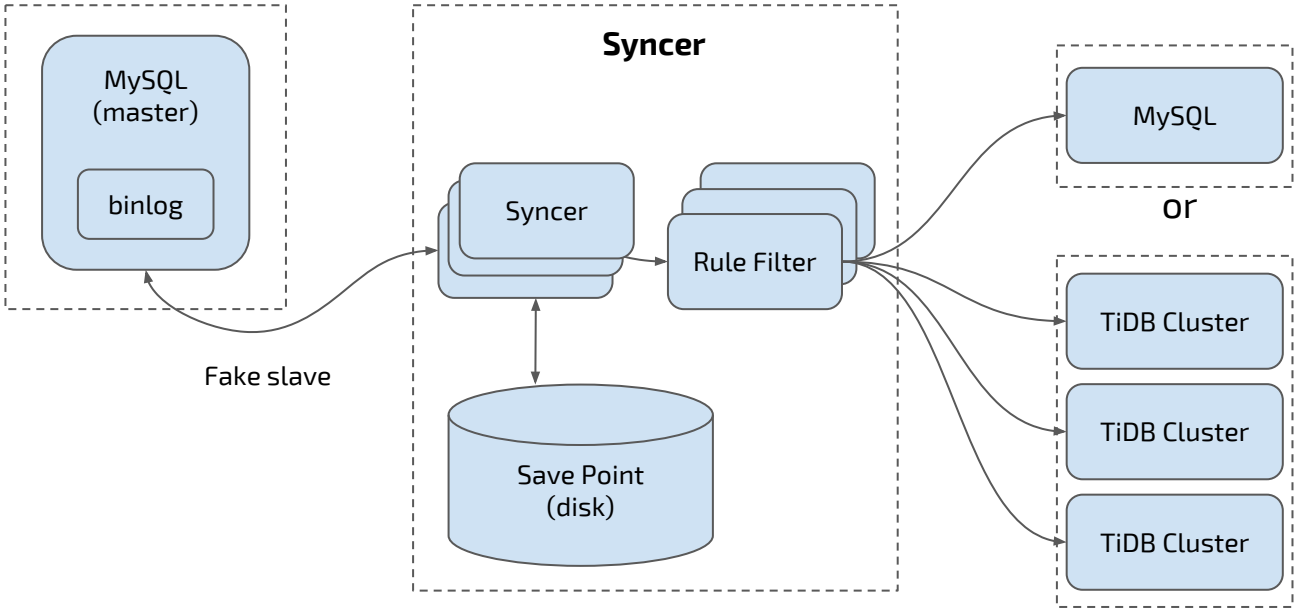
PingCAP

TiDB with the Big Data Ecosystem

PingCAP

Syncer

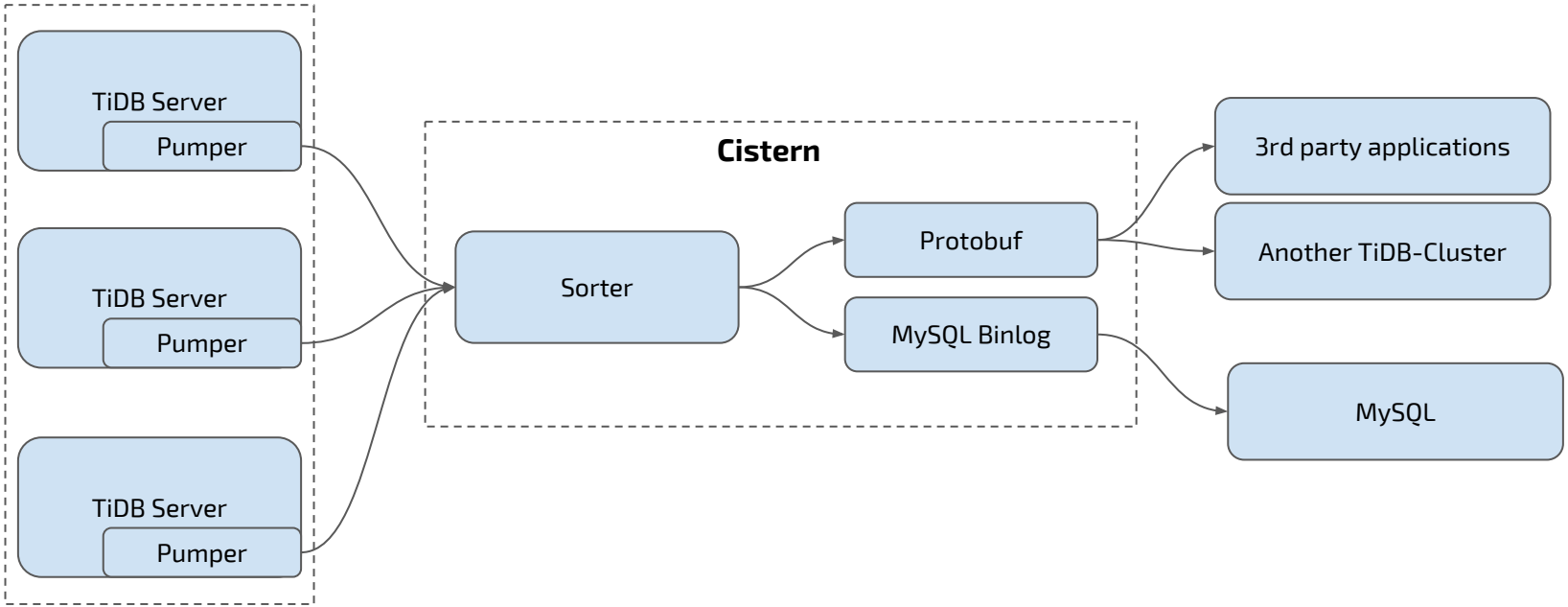
- Synchronize data from MySQL in real-time
- Hook up as a MySQL replica



PingCAP

TiDB-Binlog

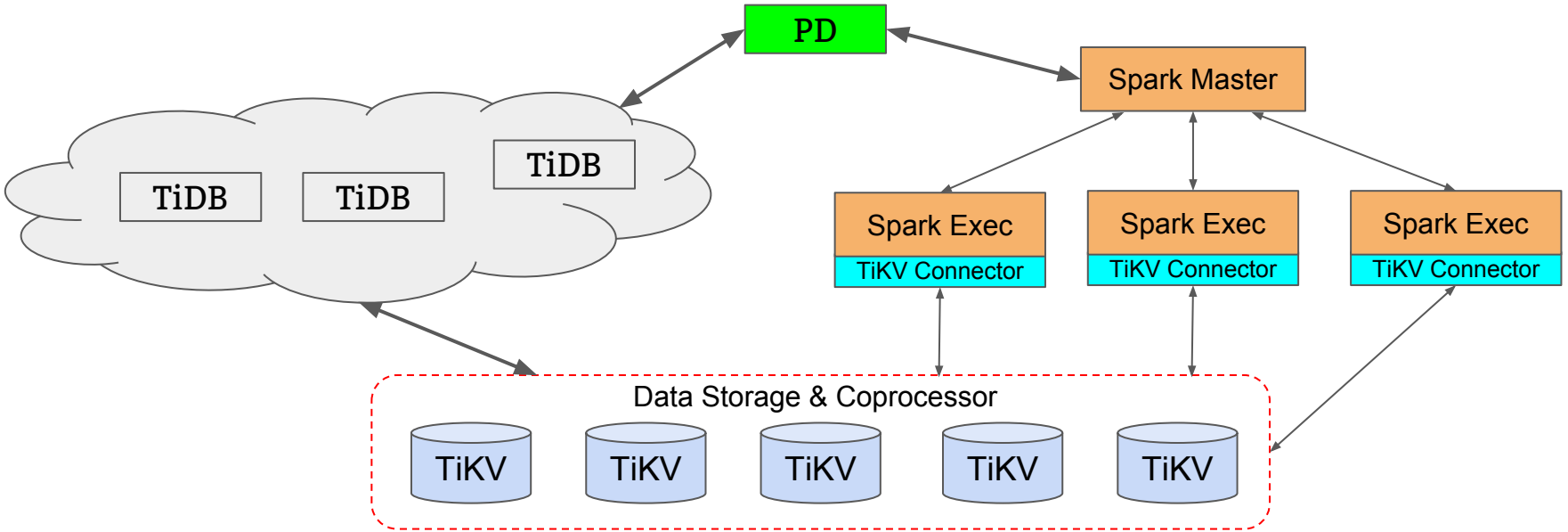
- Subscribe the incremental data from TiDB
- Output Protobuf formatted data or MySQL Binlog format(WIP)



PingCAP

TiSpark

TiDB + SparkSQL = TiSpark



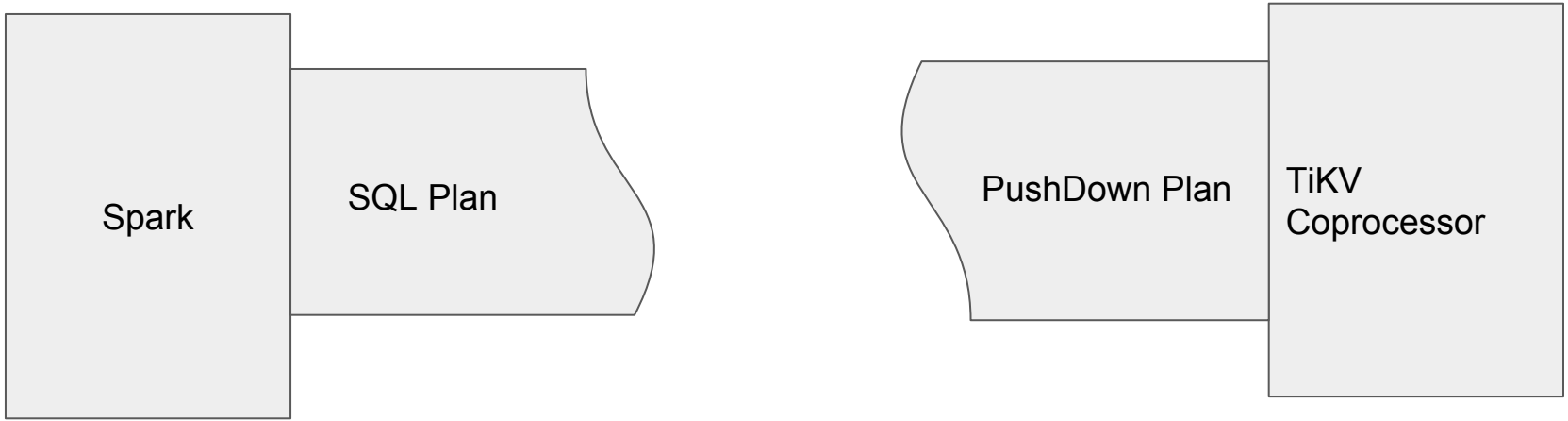
PingCAP

TiSpark

- TiKV Connector is better than JDBC connector
- Index support
- Complex Calculation Pushdown
- CBO
  - Pick up right Access Path
  - Join Reorder
- Priority & Isolation Level

PingCAP

Too Abstract? Let's get concerte.



SQL: Select sum(score) from t1 group by class
where school = "engineering";

Pushdown Plan: Sum(score), Group by class, Table:t1
Filter: School = "engineering"

PingCAP

Use Case

Use Case	MySQL	Spark	TiDB	TiSpark
Large-Aggregates	Slow or impossible if beyond scale	Well supported	Supported	Well supported
Large-joins	Slow or impossible if beyond scale	Well supported	Supported	Well supported
Point Query	Fast	Very slow on HDFS	Fast	Fast
Modification	Supported	Not possible on HDFS	Supported	Supported

Benefit

- Analytical / Transactional support all on one platform
  - No need for ETL
  - Real-time query with Spark
  - Possiblility for get rid of Hadoop
- Embrace Spark echo-system
  - Support of complex transformation and analytics with Scala / Python and R
  - Machine Learning Libraries
  - Spark Streaming

Current Status

- Phase 1: (will be released with GA)
  - Aggregates pushdown
  - Type System
  - Filter Pushdown and Access Path selection
- Phase 2: (EOY)
  - Join Reorder
  - Write

Future work of TiDB

PingCAP

Roadmap

- TiSpark: Integrate TiKV with SparkSQL
- Better optimizer (Statistic && CBO)
- Json type and document store for TiDB
  - MySQL 5.7.12+ X-Plugin
- Integrate with Kubernetes
  - Operator by CoreOS

Thanks

https://github.com/pingcap/tidb

https://github.com/pingcap/tikv

Contact me:

shenli@pingcap.com

PingCAP