Alibaba Cloud

AnalyticDB : Real-time OLAP Database System at Alibaba Cloud

Chaoqun Zhan, Maomeng Su, Chuangxian Wei, Xiaoqiang Peng, Liang Lin, Sheng Wang, Zhe Chen, Feifei Li, Yue Pan, Fang Zheng, Chengliang Chai
Alibaba Inc.

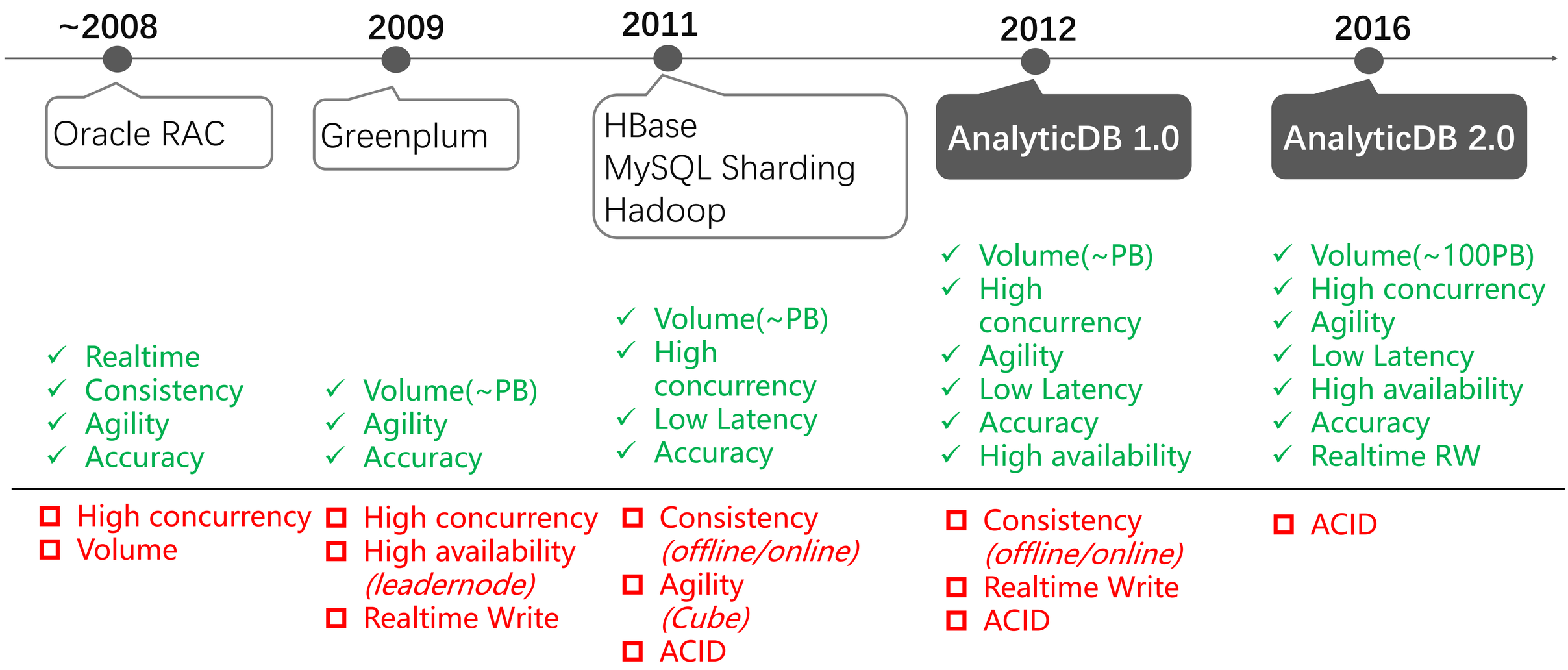
Presenter: Liang Lin, Alibaba Database BU

Outline

- Background
- Architecture
- Storage & Optimization
- Evaluation
- Future Work

1.1 Background: OLAP system evolution

Alibaba Cloud



High concurrency: ~1000 QPS(Complex Query) **Volume**:100TB+ **Realtime Write** : 10M Records/s

1.2 Background: Design Challenge

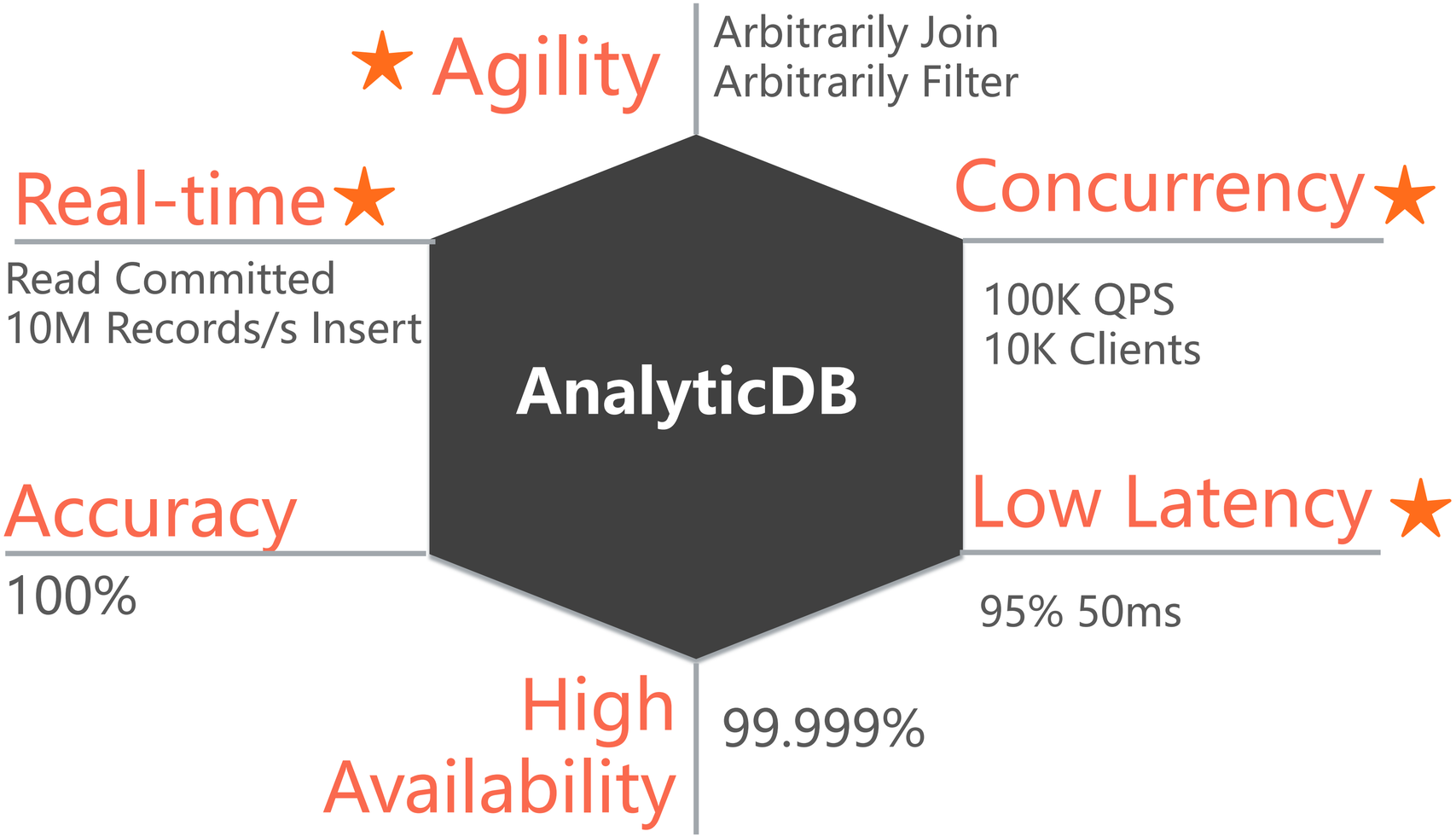
Alibaba Cloud

755M+
Active Users

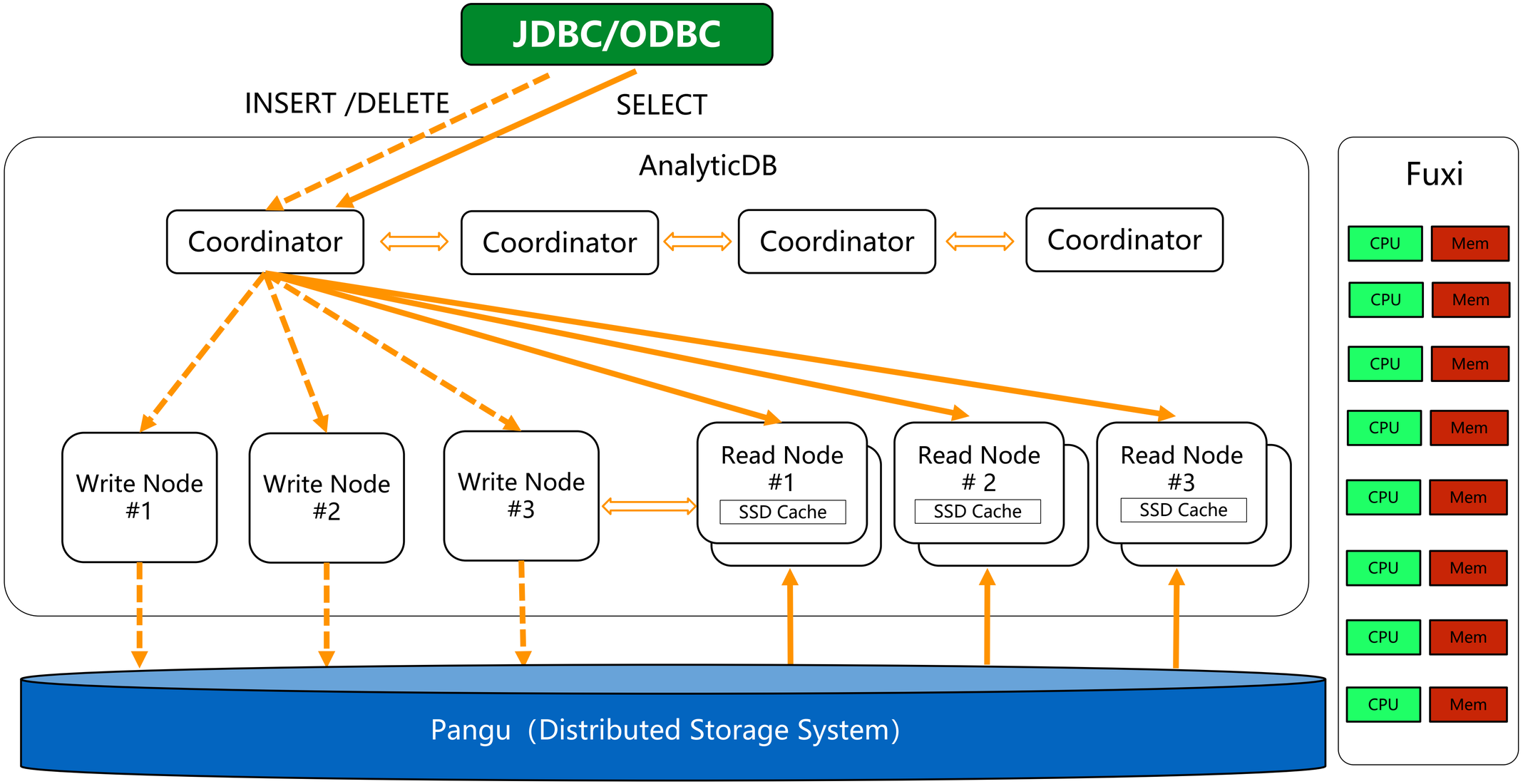
Brand
dataBank
powered by Alibaba

5+ PB
Max instance

EMS



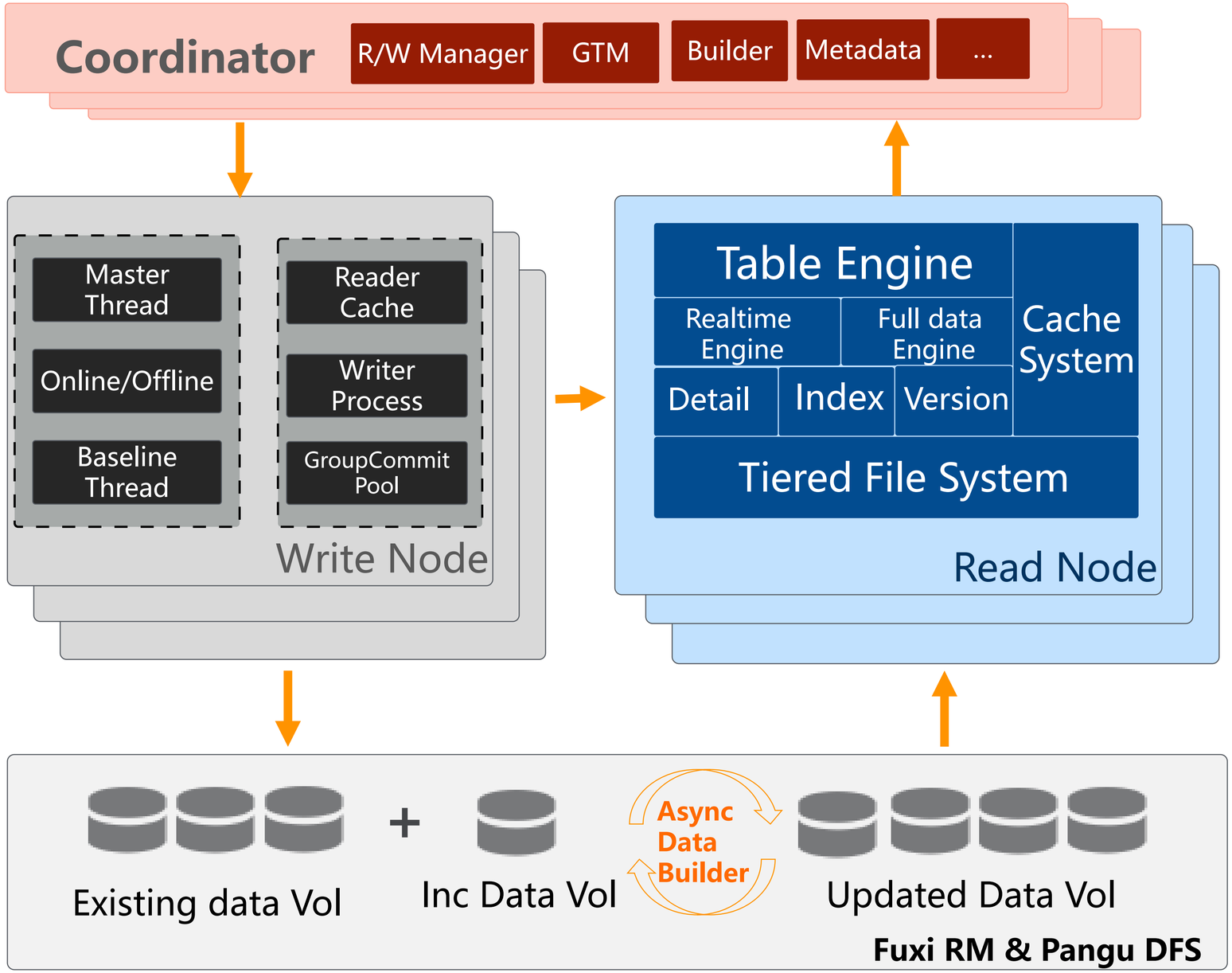
2. AnalyticDB: Architecture



3.1 Storage System Overview

Alibaba Cloud

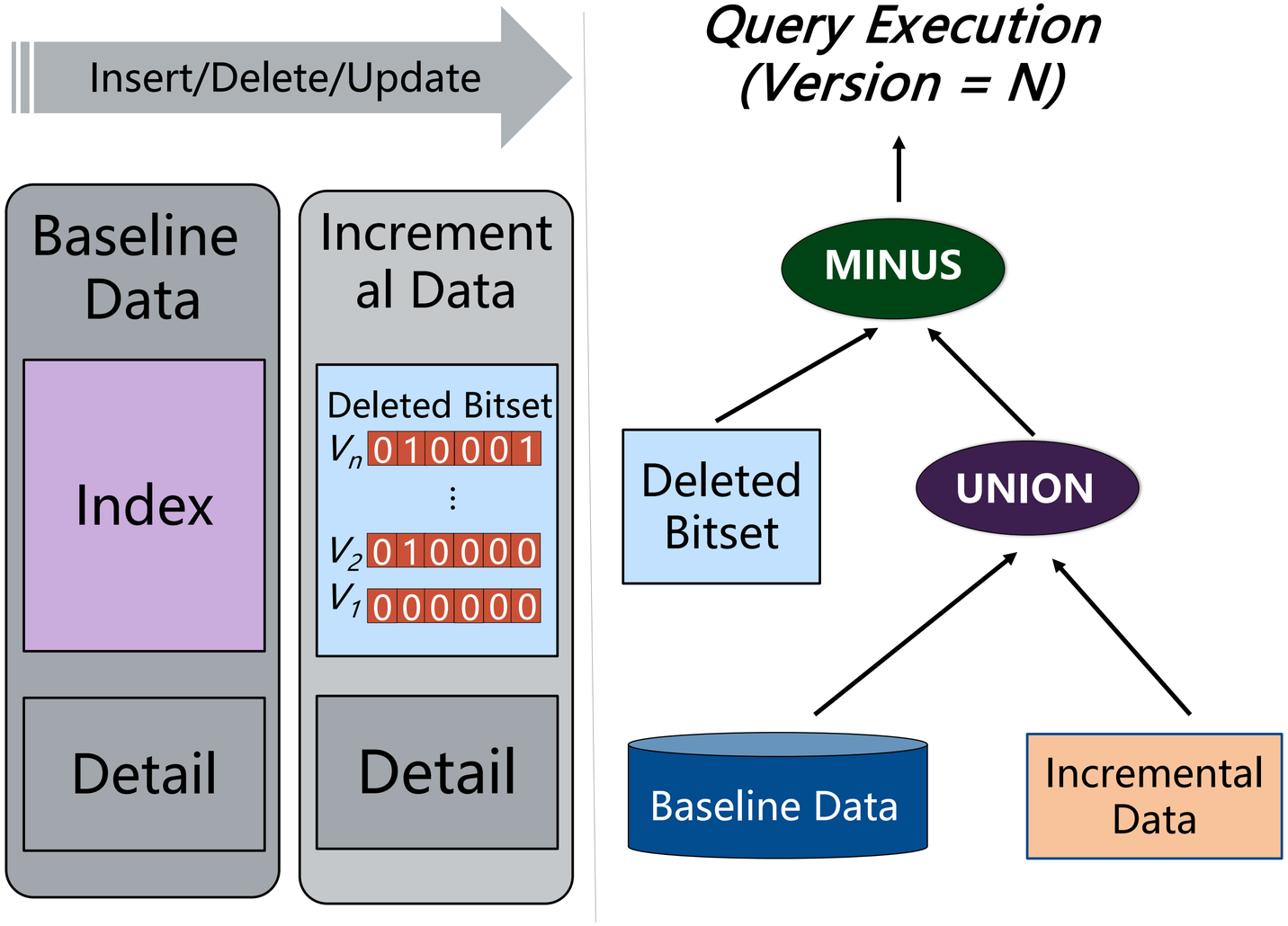
- **Read/Write decoupling**
  - High-throughput write
  - High-throughput query
- **High Scalability**
  - Scale transparently
  - up to 1024 nodes/DB
- **High Availability**
  - Fault-tolerant
  - Self-healing
  - All replicas are active
- **Strong Consistency**
  - Real-time read
- **Async Data Builder**
  - All-Column Index Builder
  - Re-partition Builder
  - Re-Clustered



3.1 Storage System: Lambda, Multi-Version

Alibaba Cloud

- **Lambda architecture**
  - Support fast insert
  - Block index for Incremental data
  - Column index for baseline data
- **Multi-Version**
  - Mark for delete with bitsets
  - Copy-on-write for dedup
  - Support snapshot read
  - Support delete and update
- **Merge**
  - Incremental index build
  - Time/size based merge
  - Merge in background
  - Data vacuum



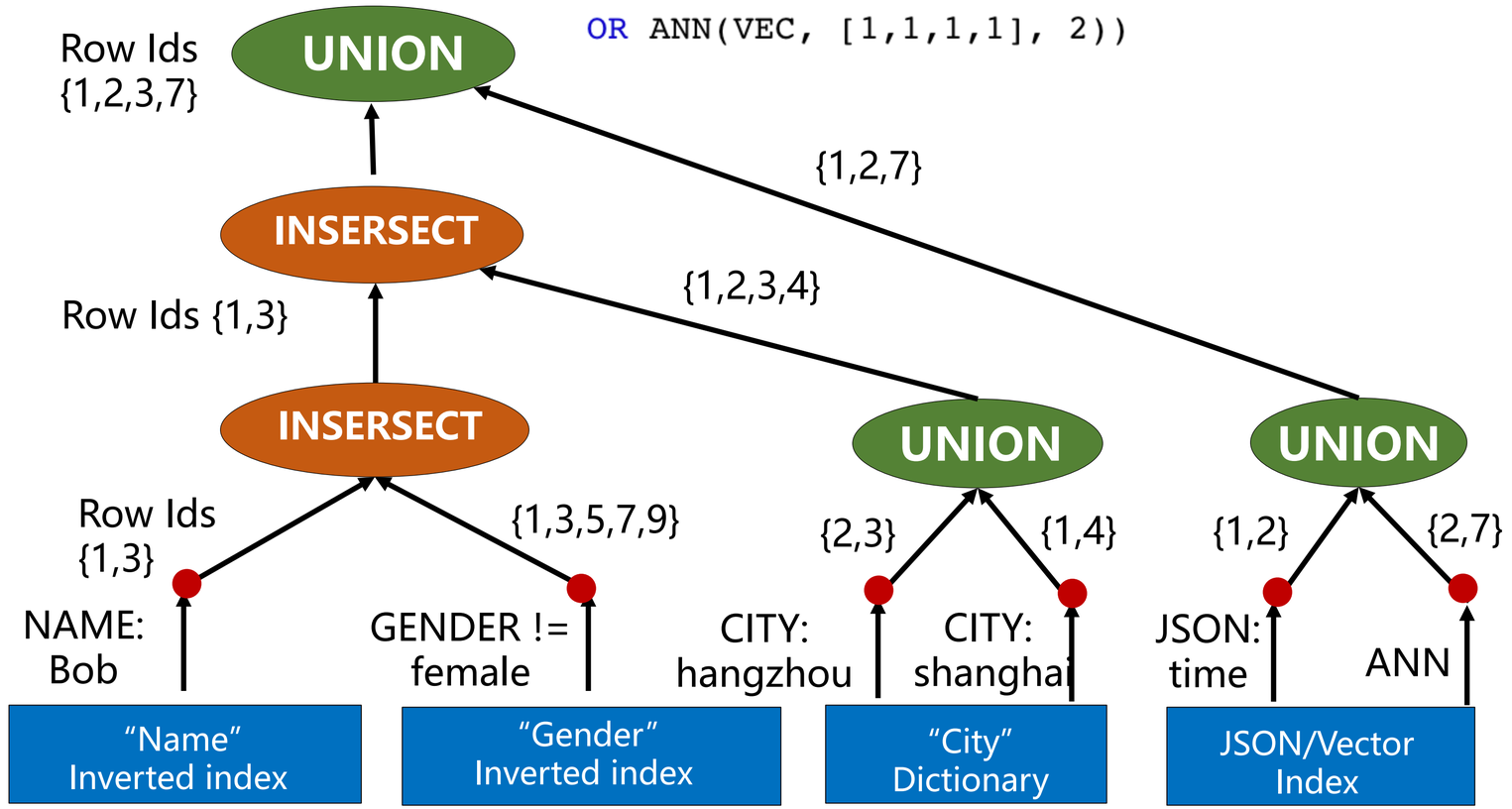
3.2 Index Computing

Alibaba Cloud

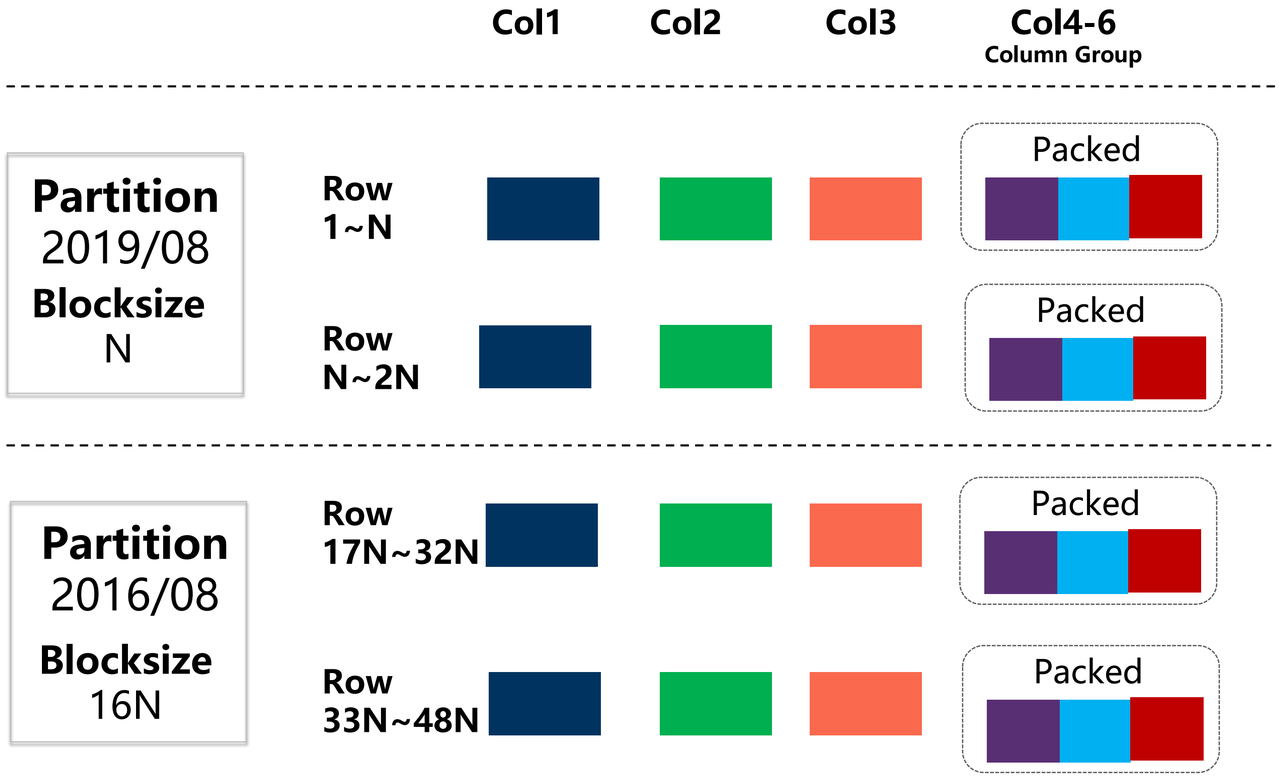
- **All-columns Indexing**
  - Indices built for all columns (automatic/optional)
  - Runtime index selection
- **High performance ad-hoc**
  - Index Computing
  - K-way merge for indexing results.
- **Various Data Type**
  - int/varchar/time/date/…
  - Full text and JSON

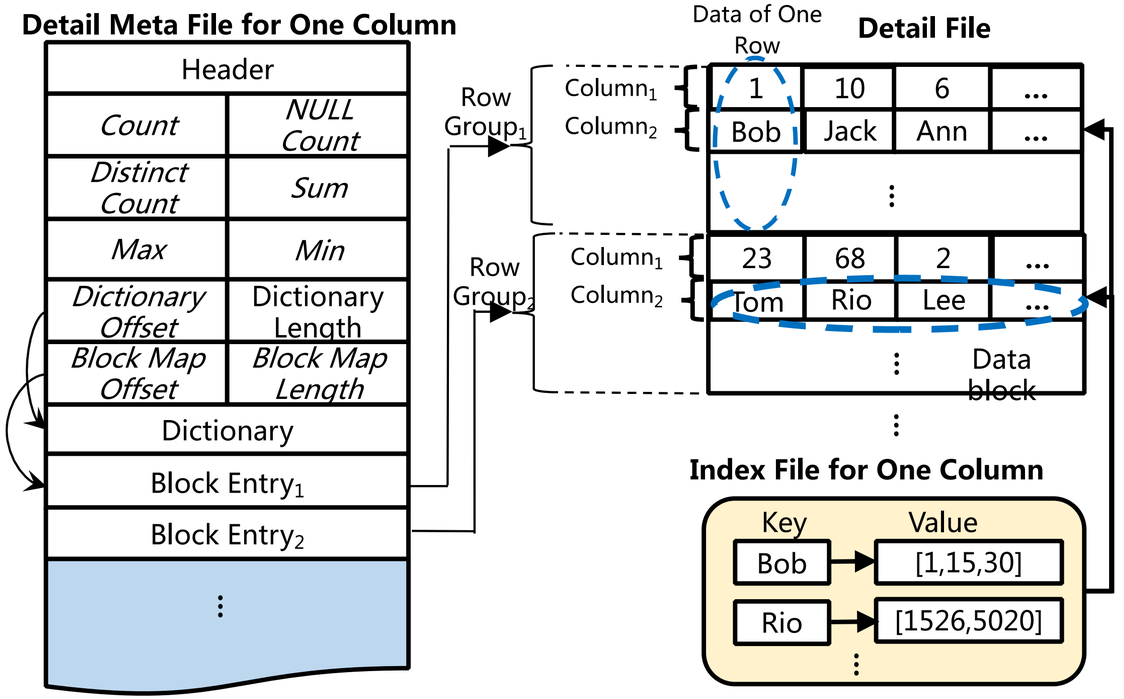
Example:

```
SELECT ... From t WHERE (name = "Bob"
    AND gender != "female"
    AND (CITY = "Hangzhou"  OR CITY = "Shanghai"))
        OR (JSON_EXTACT(ATTR,"time") > 0
        OR ANN(VEC, [1,1,1,1], 2))
```



3.3 Hybrid Row-column Storage





Multi-dimensional Analysis

- Any column join
- Complex long computing tasks, ETL

Complex Query

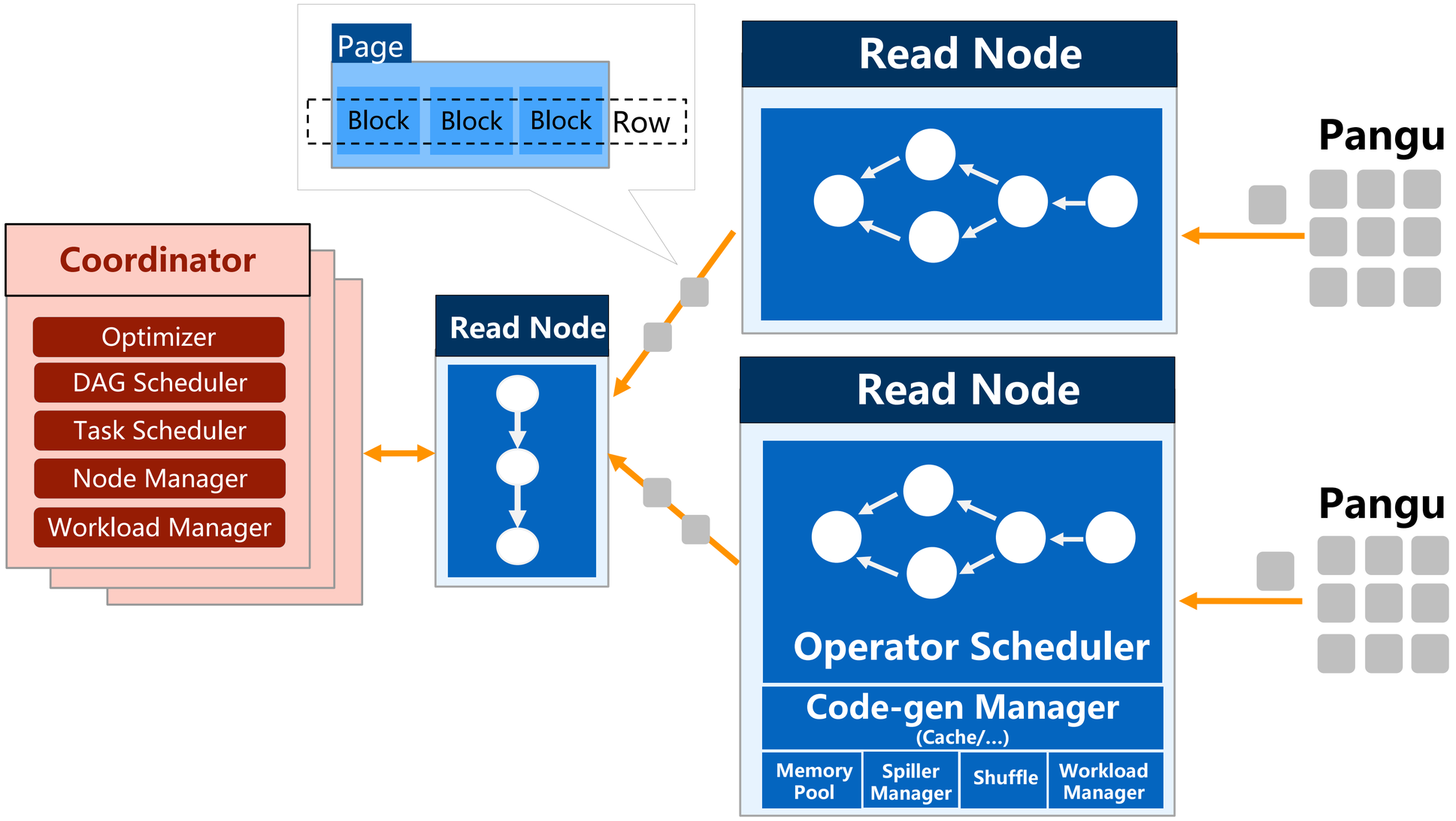
- 1000+ columns extremely wide table
- Semi-structured, large fields

Real-time Read/Write

- Live updates
- 10 million TPS
- 10K+ QPS

3.4 Execution

- Pipelining
- Codegen
- Mixed workload
- Vectorized execution
- Memory pool/cache

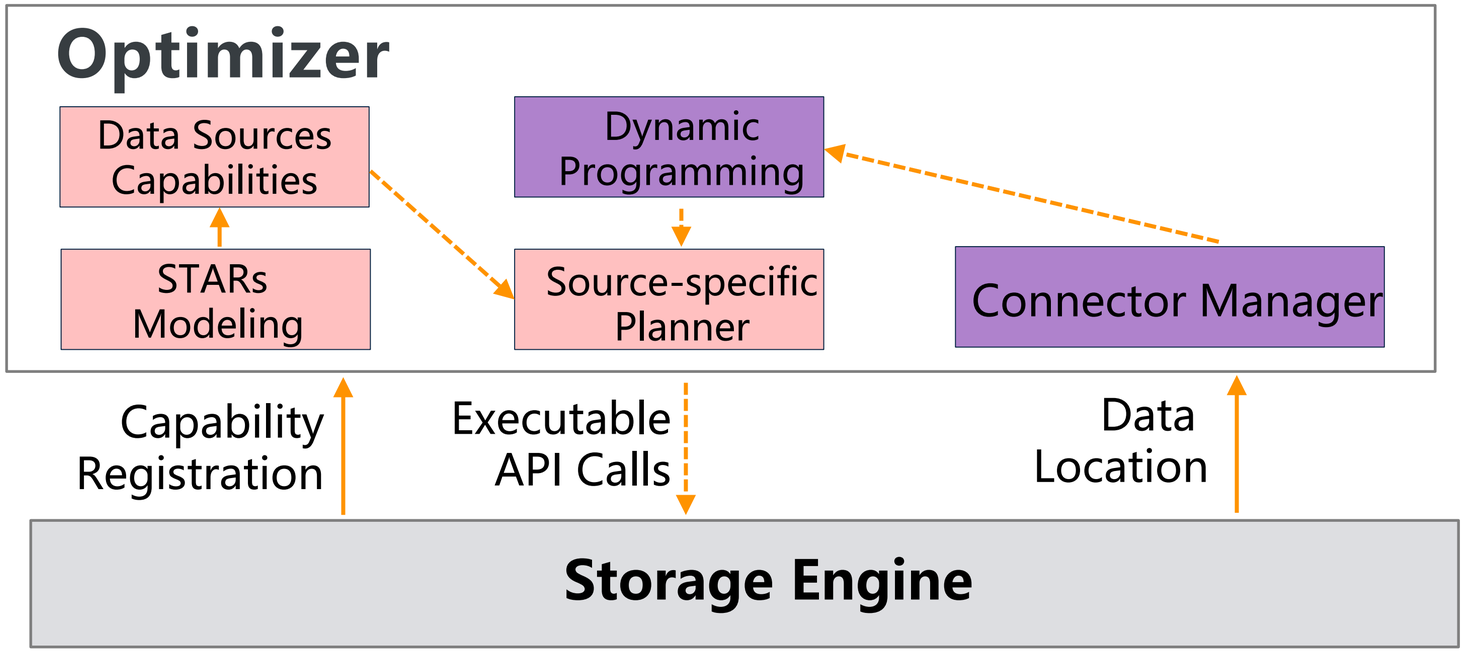


3.5 Optimization

Alibaba Cloud

Efficient Real-time Sampling

- More data meta info, better execution plan
- Index-based join and aggregation
- Less data read, less computation

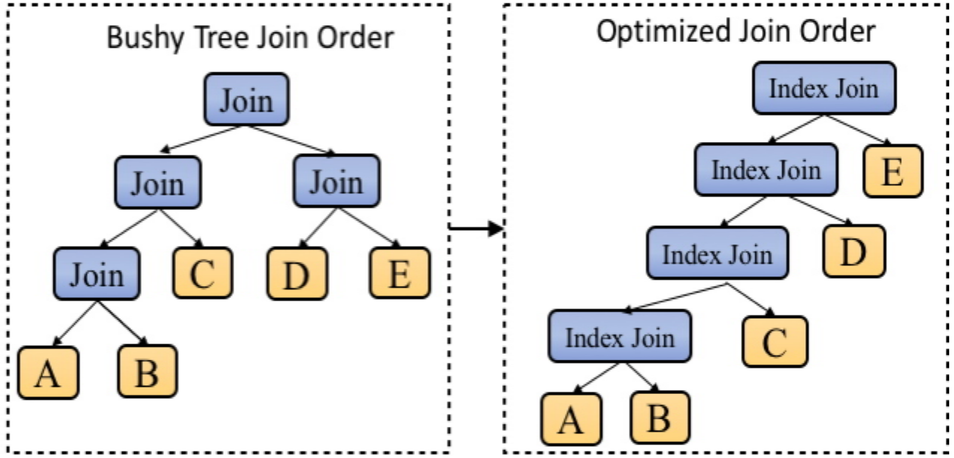


Join push-down

Predicate push-down

Join Reordering

```
SELECT …
FROM   a
       JOIN b
         ON a.id = b.id
       JOIN c
         ON b.id = c.id
       JOIN d
         ON c.id = d.id
       JOIN e
         ON d.id = e.id
WHERE …
```



4. Experiment : Setup

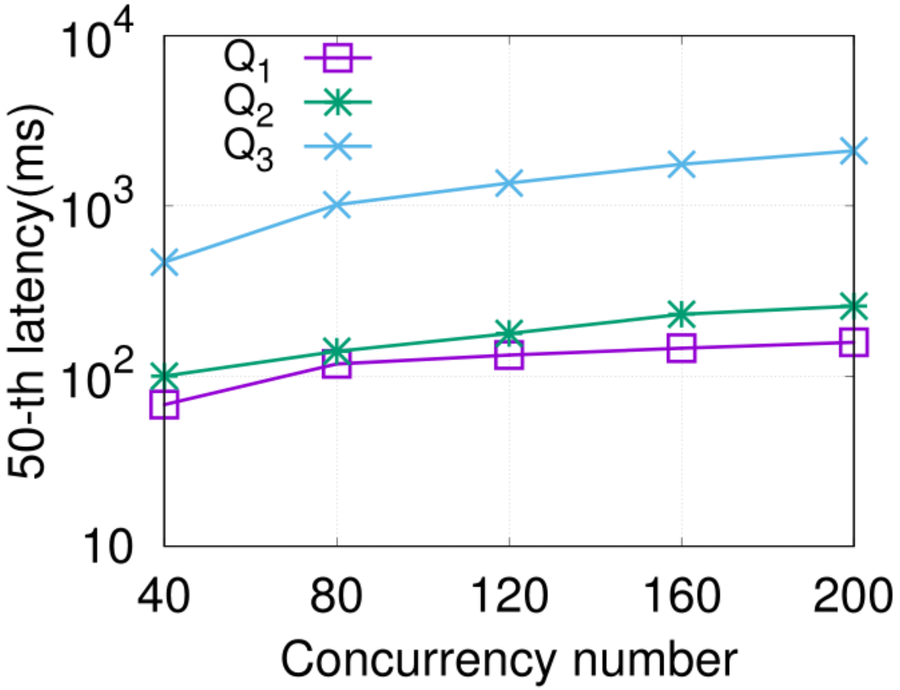
Alibaba Cloud

Query Type	Query
Full Scan (Q1)	SELECT * FROM orders ORDER BY o trade time LIMIT 10
Point Lookup (Q2)	SELECT * FROM orders WHERE o trade time BETWEEN '2018-11-13 15:15:21' AND '2018-11-13 16:15:21' AND o trade prize BETWEEN 50 AND 60 AND o seller id=9999 LIMIT 1000
Multi-table Join (Q3)	SELECT o seller id, SUM(o trade prize) AS c FROM orders JOIN user ON orders.o user id = user.u id WHERE u age=10 AND o trade time BETWEEN '2018-11-13 15:15:21' AND '2018-11-13 16:15:21' GROUP BY o seller id ORDER BY c DESC LIMIT 10;

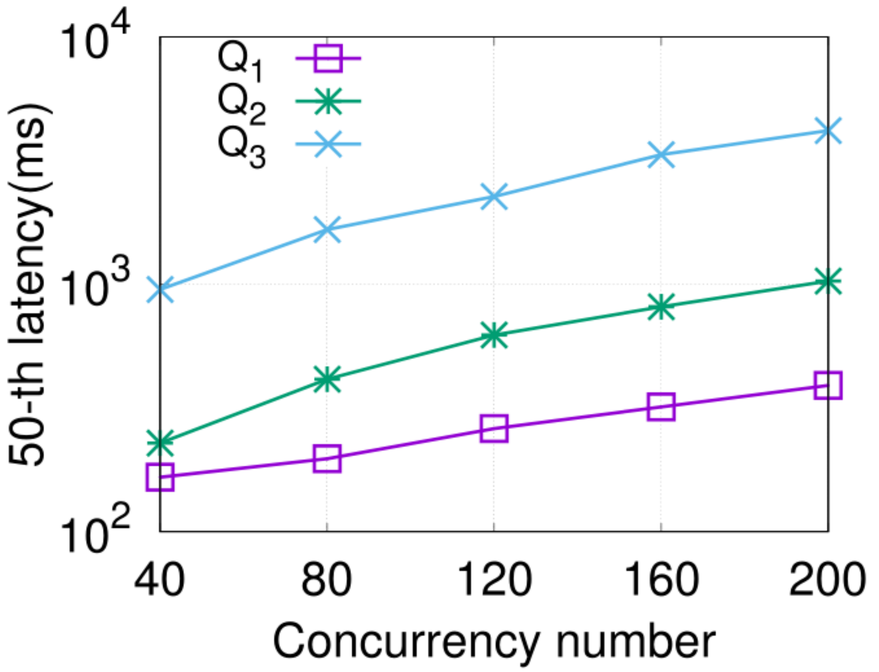
- ✓Eight Physical Machines
  - ✓Intel Xeon Platinum 8163 CPU, @ 2.5 GHz
  - ✓300GB main memory and 3TB SSD
  - ✓10Gbps Ethernet network
- ✓Deployment of AnalyticDB
  - ✓4 coordinators
  - ✓4 write nodes, and 32 read nodes
- ✓Workloads
  - ✓1TB and 10 TB

4. Experiment : Latency Analysis

Alibaba Cloud



1TB

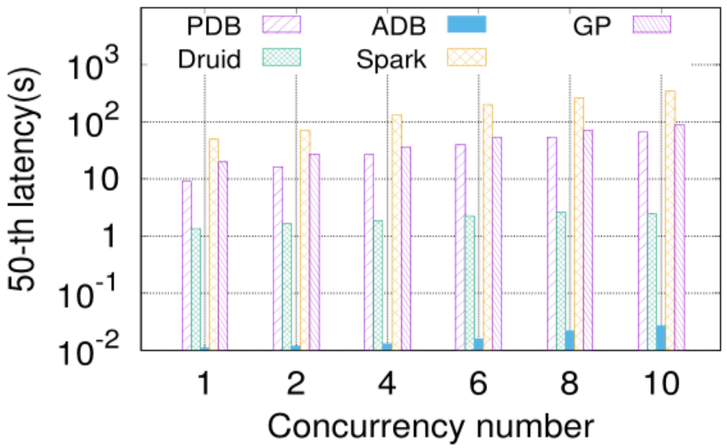


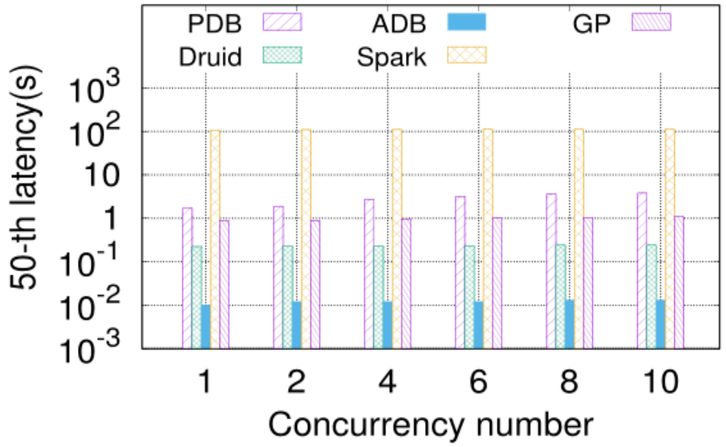
10TB

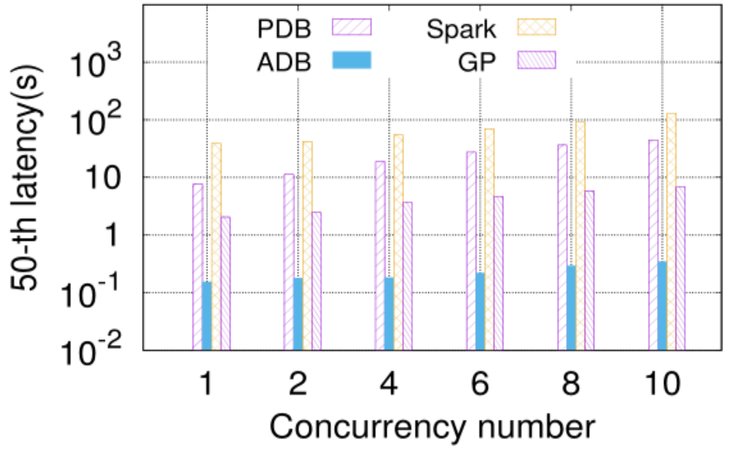
- ✓ All three types of queries are completed within seconds
- ✓ Query performance of AnalyticDB is slightly impacted by the table size

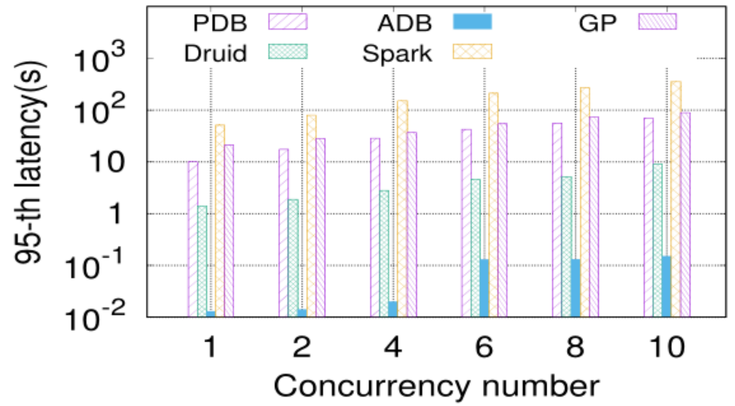
4. Experiment : Latency Comparison

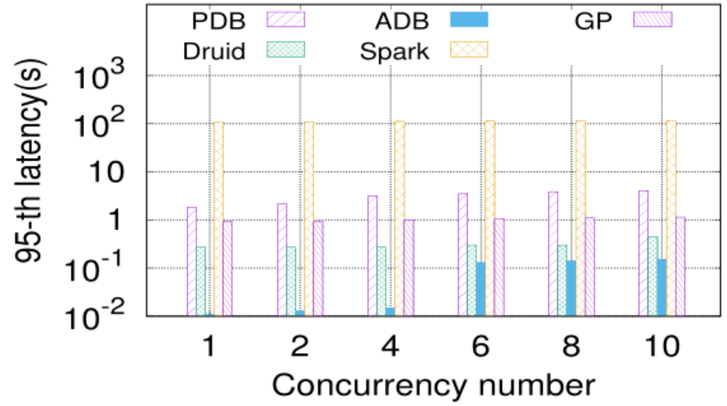
Alibaba Cloud

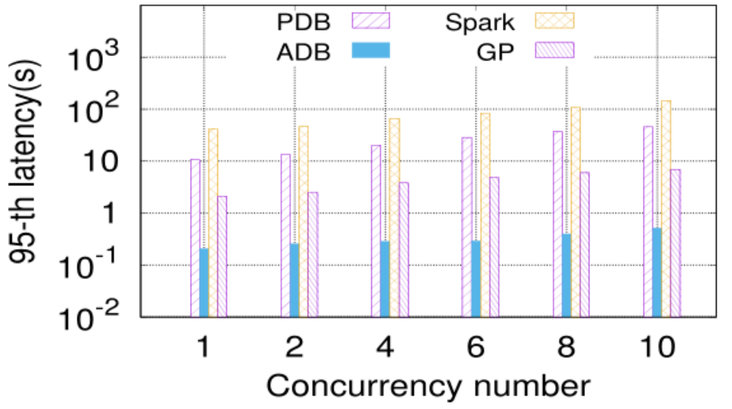










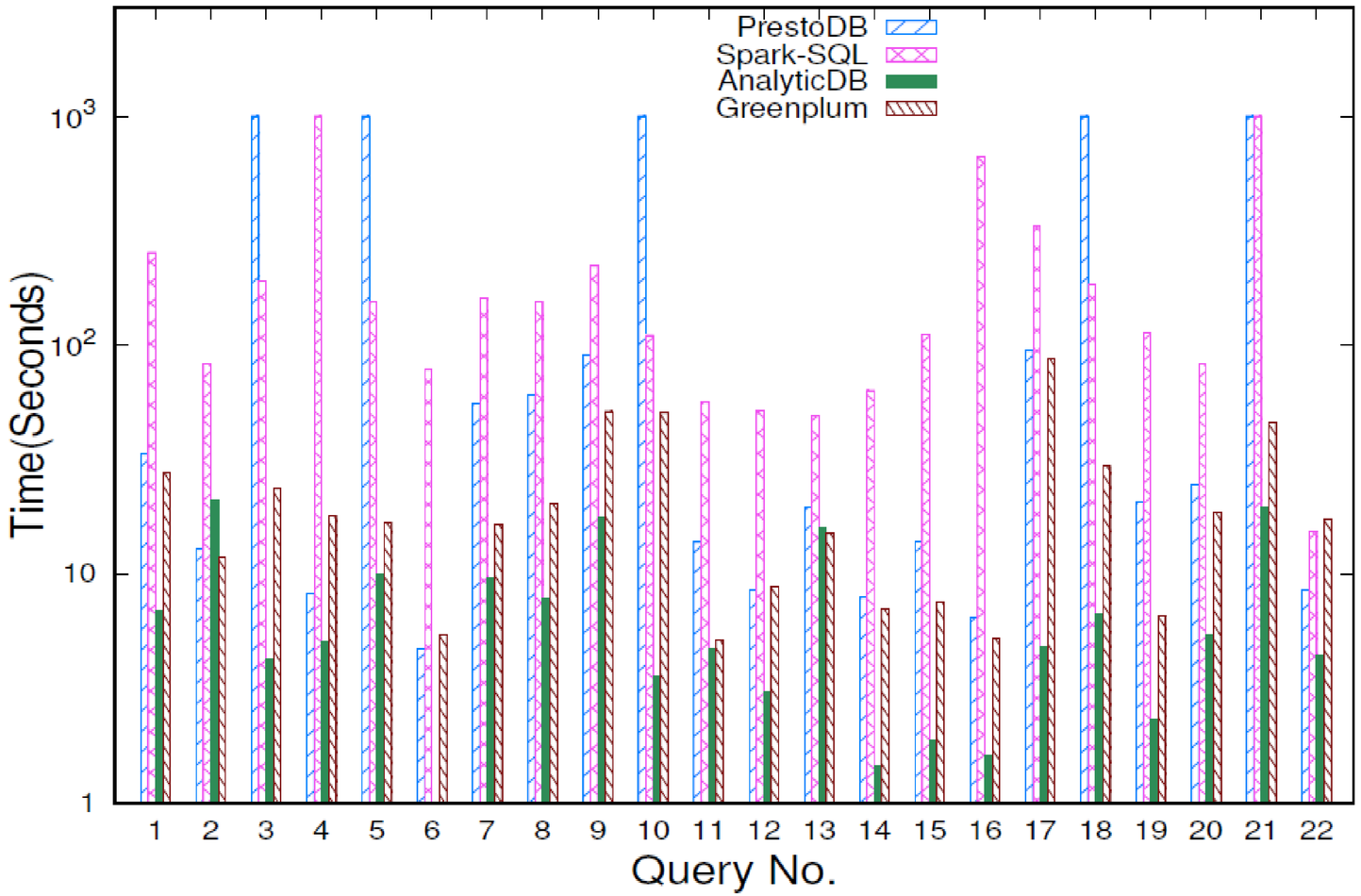


4. Experiment: TPCH Evaluation

Alibaba Cloud

Advantages:

- ✓ Pipeline-process
- ✓ All-column index
- ✓ Hybrid row-column storage
- ✓ Runtime cost-based index path selection
- ✓ K-ways merging and composite predicates pushdown
- ✓ Vectorized execution engine and optimized CodeGen



5. Future Work

Alibaba Cloud

- **Separation storage/computation**

- **HTAP**
- **Multi-Master**
- Cross AZ HA

- **Serverless**
- **Separation storage/computation/ memory**
- ML Optimizer

Alibaba Cloud

Q & A