NebulaGraph | VELODB

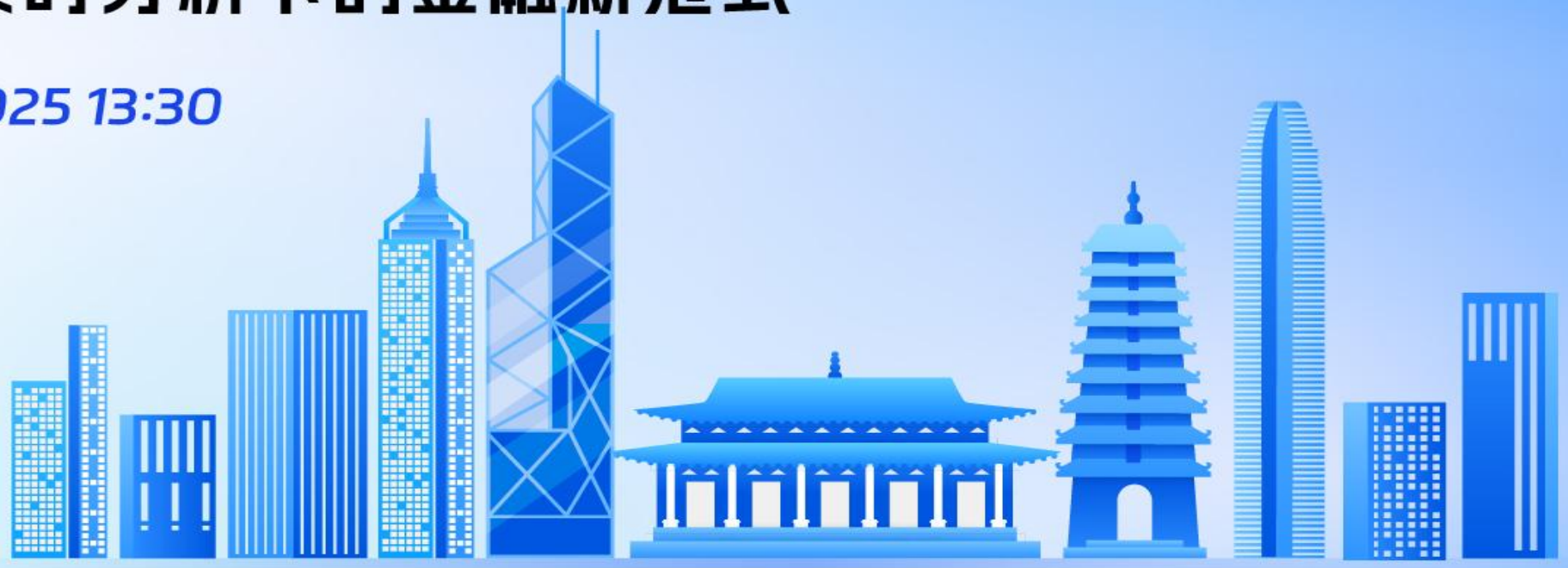
Deep Relationships & Real-Time Analytics:

Architecting Modern Financial Data Stacks

深度关联与实时分析下的金融新范式

August 22, 2025 13:30

Hong Kong



Graph Empowers Financial Institutions with Penetration-Based Management

----By HU XIN FROM CSCS

NebulaGraph

中证数智 China Securities CreditTech Services

CONTENTS

01 About us

02 Knowledge Graph Construction

03 Graph Algorithms——Capital Group Identification

04 Knowledge Graph Application Scenarios

China Securities Credit Investment

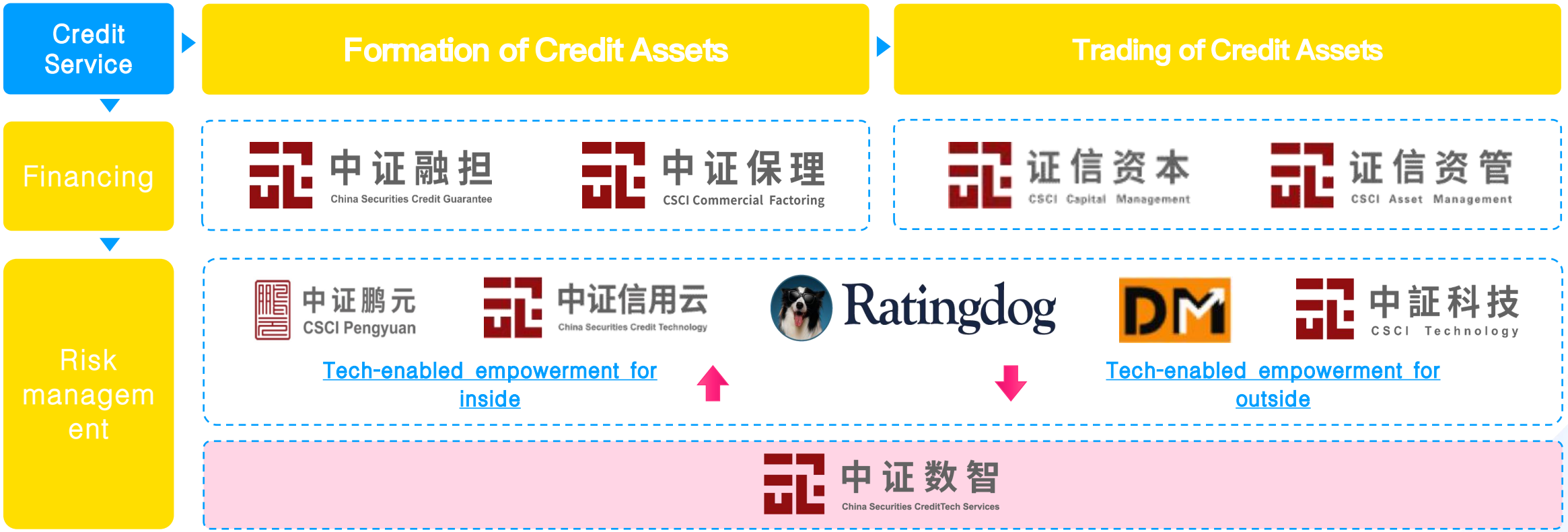
China Securities Credit Investment Co., Ltd. (“China Securities Credit Investment”) is a leading Chinese credit-tech service provider, committed to building infrastructure for the whole life cycle management of credit assets through technology-driven means, provides full credit value chain services including credit risk management, credit enhancement, and credit asset transaction management for institutional clients.

Brief Profile

- **Establishment date**：May 27, 2015
- **Registered capital**: 4,585,980,000 RMB
- **Shareholder background:** Jointly funded by 35 shareholders with relatively dispersed equity ratios. The top five shareholders are Soochow Securities (4.91%), New Horizon Capital (4.86%), Guoyuan Securities (4.36%), Qianhai Financial Holdings (4.36%), and Essence Securities (4.36%)
- **Employee size:** over1,000 employees (60%+ in technology/risk mgmt)
- **Rating grade:** AAA
- **Honors**：KPMG "China's Leading Fintech 50 Enterprises" (2019–2022) and the first batch of comprehensive credit service institutions of the National Development and Reform Commission

China Securities Credit Investment

China Securities Credit Investment owns 8 majority-owned tier-1 subsidiaries, which include China Securities CreditTech Services Co., Ltd., CSCI Pengyuan Credit Rating Co., Ltd., China Securities Credit Technology Co., Ltd., China Securities Credit Guarantee Co., Ltd., CSCI Commercial Factoring Co., Ltd., CSCI Technology Co., Limited,CSCI Asset Management Co., Ltd. and CSCI Capital Management Co., Ltd.



China Securities CreditTech Services -Our Team

Distribution


Shenzhen *70+*


Shanghai *70+*


Beijing *40+*


wuhan *100+*

- Data Specialists：100+
- R&D：65+
- Delivery：70+
- Solutions：30+

Diversed Talent

- Finance and IT：60%
- Executive Team experience：25+ years
- Retention：60% with ⩾3 years tenure
- Mid Managers：15+ years

China Securities CreditTech Services Co., Ltd. -- 中证数智科技（深圳）有限公司

China Securities CreditTech Services Co., Ltd. (hereinafter referred to as “CSCS") was established on September 28, 2015 with a paid-up capital of RMB 215.45 million.

CSCS is committed to using cutting-edge technology to help various financial institutions improve their credit risk identification, measurement, monitoring and management capabilities, and provide intelligent credit data, credit evaluation, correlation graph application, intelligent early warning, regulatory technology and other services for regulators, investment institutions, intermediaries and other financial market participants.

Enabling the digital transformation of credit management

credit data

credit data

Complete low-frequency and high-frequency data, product data, and macro- and meso-data of core companies in the capital market

credit evaluation

The Credit Asset Internal Assessment System (CIAS)

Integrated "data-model-system" credit risk management solution covering bond, non-standard, ABS and credit products

Client risk management Platform

Intelligent Client Relationship Graph

Based on the Enterprises’ profile and relationship Graph data of the whole registered Enterprises, we can build a client platform for institutions to conduct investigations and analyses on enterprises

Group client risk management

Group-Client risk management in Credit risk,including Concentration management, Credit facility assesment & Risk Transmission

Intelligent warning

Intelligent Early Warning System

Realizing the comprehensive portrayal of enterprise early warning information and the efficient monitoring of risk variations

RegTech

Regulatory service systems

Provide data support, analysis system, application system and other related services for regulatory agencies to help them improve their efficiency

Honors and Qualifications

- ✓ national high-tech enterprises
- ✓ "Specialized, specialized and new" enterprises
- ✓ "CMMI Software Capability Maturity Level 3" certification
- ✓ Qianhai Excellent Financial Case - CreditMaster
  investor suitability
  Forensics
  CSI Cloud Chart
- ✓ "ISO9001 Quality Management System" certification
- ✓ Data products listed on Shanghai Data Exchange and Zhengzhou Data Exchange (four each)

NebulaGraph

中证数智 China Securities CreditTech Services

CONTENTS

01 About us

02 Knowledge Graph Construction

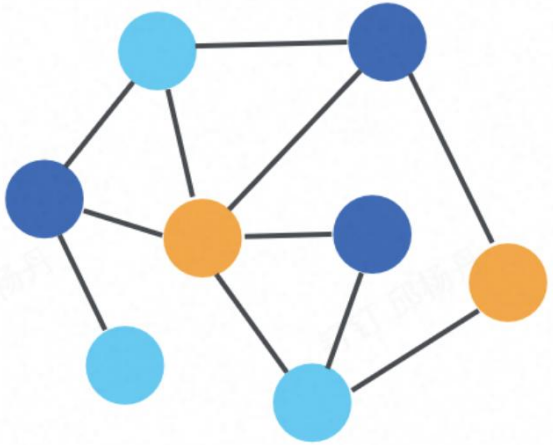
03 Graph Algorithms——Capital Group Identification

04 Knowledge Graph Application Scenarios

What is Graph Database

Graph database is a database that stores entities of related data as **vertex** and the relationships between data as **edge**.

NebulaGraph is a highly performant linearly scalable graph database available for use via a shared-nothing distributed model.



- Use table to store data
- Query via SQL
- Slow performance and high resource consumption

✓ Use graph to store data
✓ Query via ISO-GQL
✓ Built for super large-scale graphs with milliseconds of latency

NebulaGraph

中证数智

CSCS Data Asset Encyclopedia

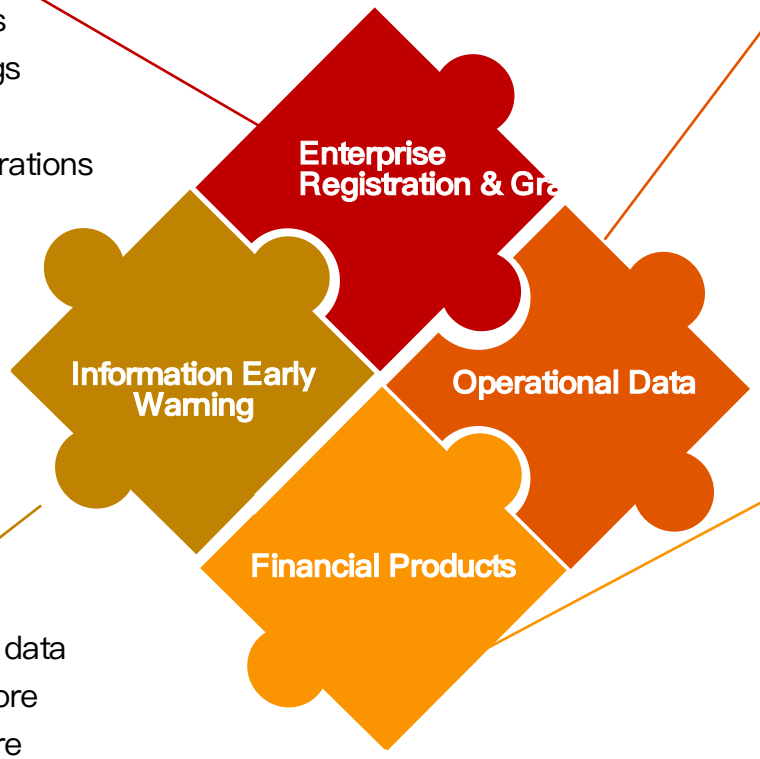
Micro

Enterprise Registration & Graph

- Business registration
- Shareholders /Executives
- Branches
- Investment
- Business changes
- related parties
- Annual reports
- External ratings
- Mortgage
- Abnormal operations
- Administrative penalties
- Graph data
- ……

Information Early Warning

- News &Sentiment
- Announcement
- Judicial data
- Integrity data
- Early warning data
- Sentiment Score
- Warning Score
- ……



Operational Data

- Enterprise exposure
- Regional data
- Financial data
- Financial indicators
- Operational indicators
- Guarantee information
- ……

Financial Products

- Bond information
- Stock information
- Private equity funds
- ABS information
- REITS
- ……

Meso&Macro

China Macro

- Industry economy
- Industry prosperity
- National economic accounting
- Population
- Policy data
- ……

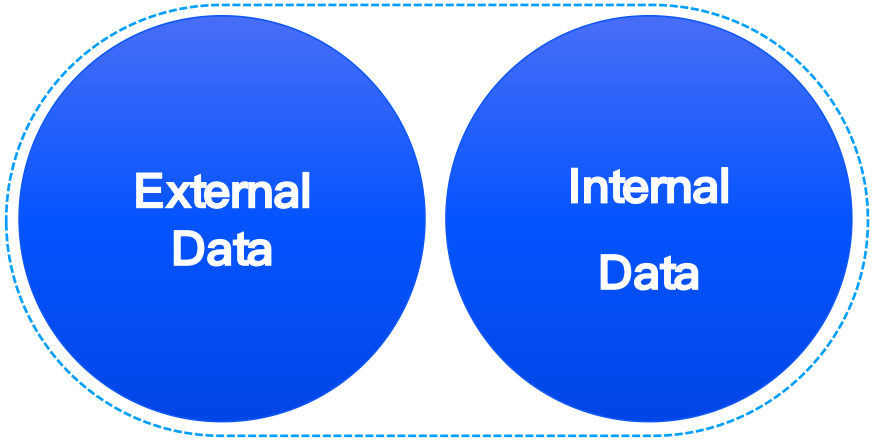
Regional Data

- Regional economy
- Regional finance
- Regional debt

Basic Data Sources for Financial Institution Knowledge Graph

Domestic & Oversea data: enterprises, products & personal data

Relational database and graph database



Integrate non - public relationship information obtained from business due diligence

Some external relationship information is lagging, replace with updated information obtained by account managers

Load internal business relationships

The enterprise graph construction process

Traditional databases → Exporting a CSV file → With the Graph Update tool → graph databases

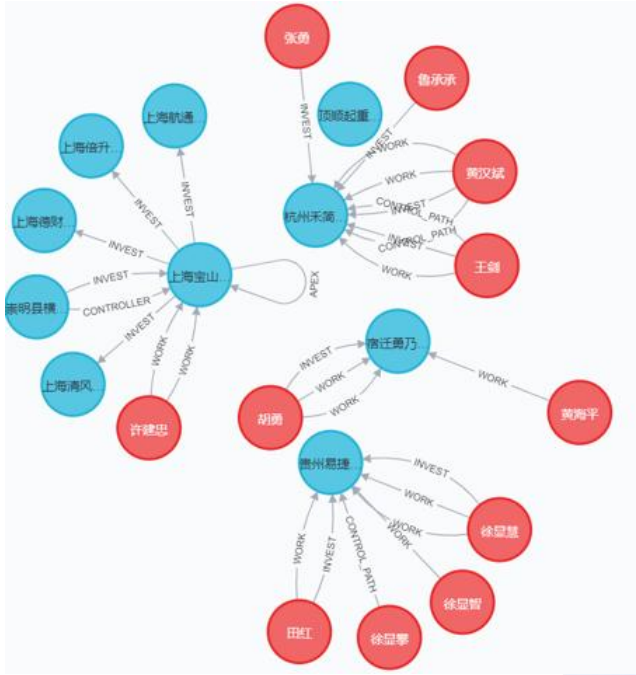
- Company info data
- Company shareholder data
- Company executive dta

→

- Node file
- Node properties file
- Relationship files
- Relationship properties file

→

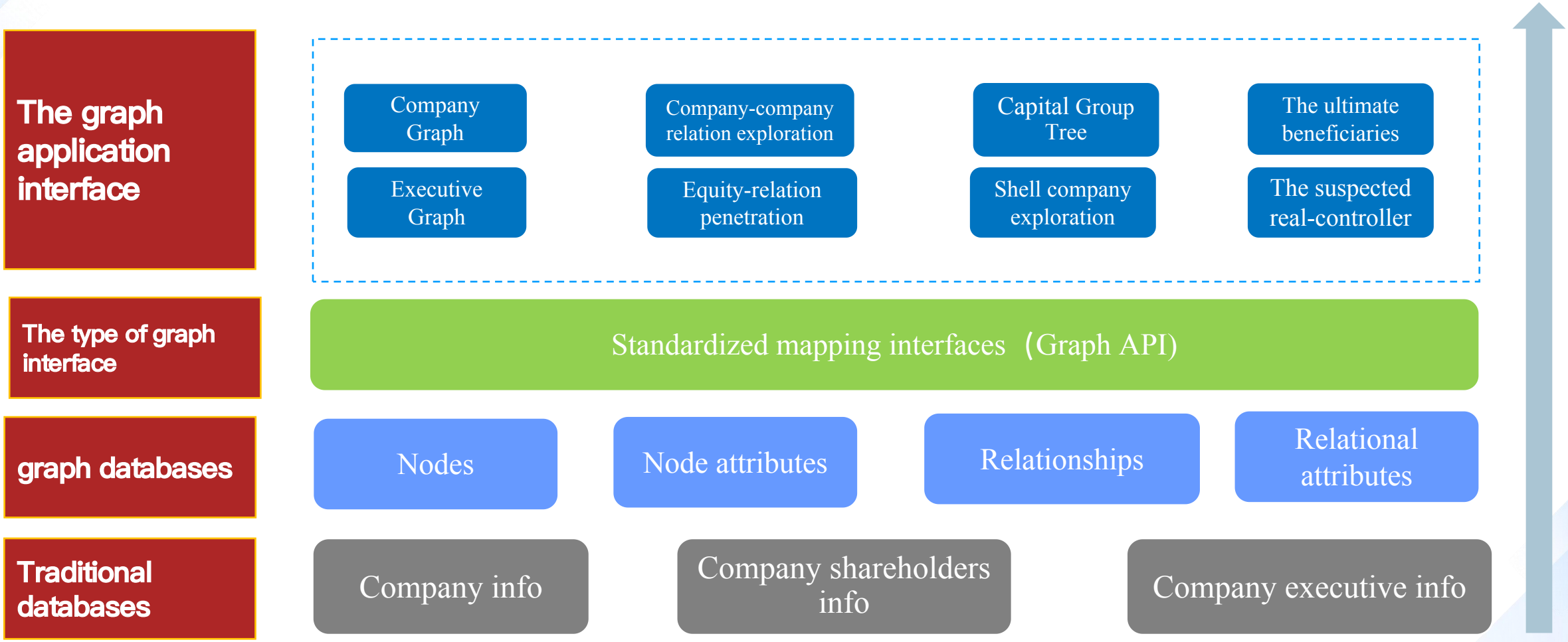
→



NebulaGraph

中证数智

Enterprise graph construction methodology



Overview of Financial Institution Graph Construction

- Constructing relationship graphs for enterprises across the entire market, with a cumulative coverage of 300 million + nodes and 400 million + relationships.
- Mature experience in graph data generation, updating, integration, and application, leveraging graph databases to provide continuous operation, maintenance, and mining application services tailored for financial scenarios.

Node Types

Corporate Entities: 120 M

Public Person Records: 140 M

Financial Products: 2

Relationship Types

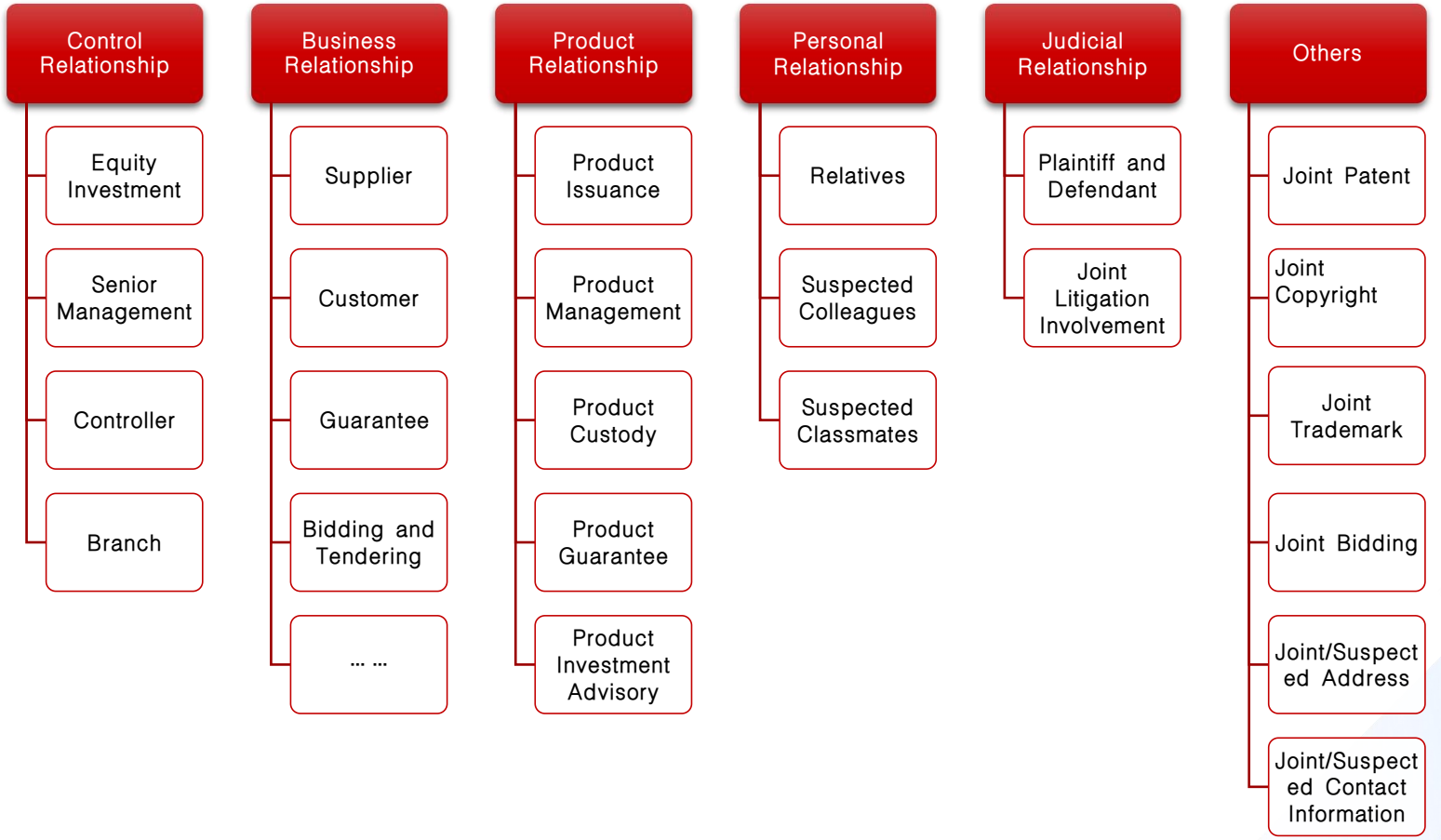
Control Relationships: 300 M pairs

Business Relationships: Millions

Financial-Product Relationships: 3 M pairs

Personal Relationships: ~1 M pairs

Litigation Relationships: ~6 M pairs



NebulaGraph

中证数智

Big Data and Graph Technology empowering diverse application Scenarios

"Rules" Perspective

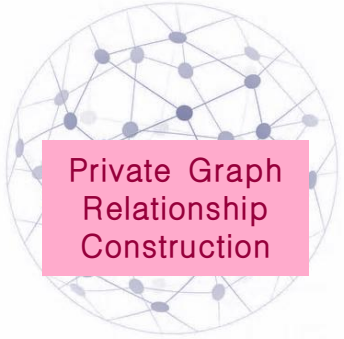
Business rules based on expert experience and historical data accumulation.
Implement rules using "Big Data", "Graph Technology", and "AI Algorithms" to enhance rule **efficiency, coverage, and accuracy.**

"Intelligence"Perspective

Discovering unknown risks and patterns relying on machine learning, graph computing, etc.

Discover unknown risks and patterns using "Big Data", "Graph Technology", and "AI Algorithms".

Enterprise-Level Big Data & Graph Platform



Financial Institution Business Scenarios

Front Office

Client Management

- Customer Due Diligence
- Client profiling
- Opportunity explore
- ... ...

Middle & Back Office

Risk Management

- Suspected relationship explore
- Centralized management
- Shell Company Identification
- Risk Contagion
- Group Client Management
- Risk Investigation
- Risk Label Models
- ... ...

+

Compliance Management

- Related parties & related-party transactions
- AML Suspicious Transaction
- AMLBeneficial Owner
- Abnormal Transaction Models
- Conflict of Interest Check
- ... ...

NebulaGraph

中证数智

CONTENTS

01 About us

02 Knowledge Graph Construction

03 Graph Algorithms——Capital Group Identification

04 Knowledge Graph Application Scenarios

Capital Group Tree

INPUT

- Equity relationships of all companies

Calculation

Controller Calculation

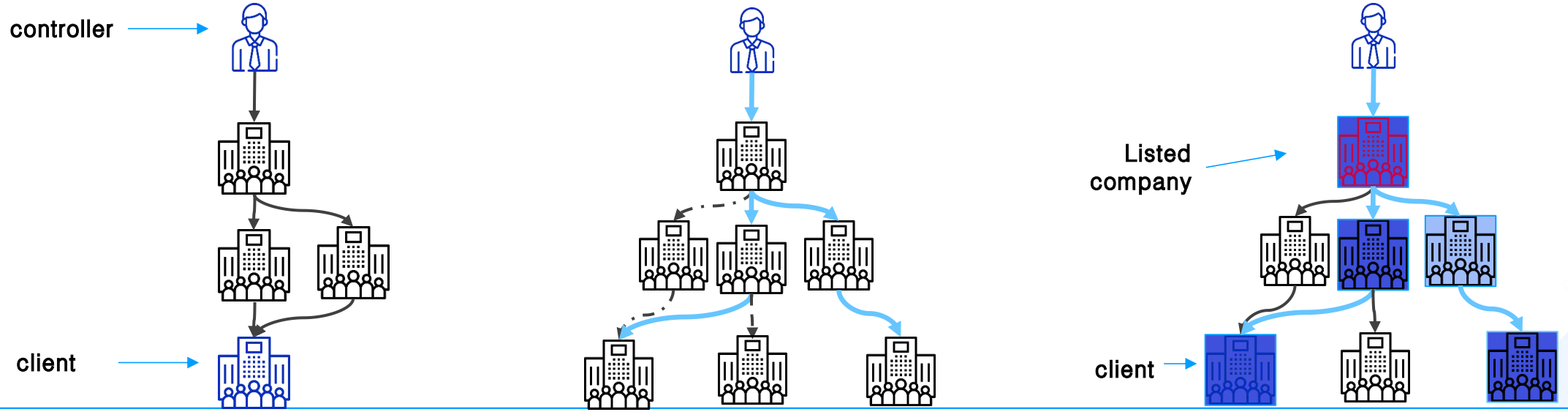
- Suspected real controller
- The largest voting shareholder
- Special adjustments：governments revolving holding etc.

Minimum real-control path

- The path to real-controller
- Elimination non-actual control paths
- Remain real-control relationships

Traversing all the members

- breadth-first or depth-first
- Companies merged with same controller
- Constructing Factions of Big group companies Sub-Branch Groups



OUTPUT

- Weekly calculations are updated to identify group relationships: name and IDs, corresponding to all group top names and IDs, control paths, shareholding percentages, etc

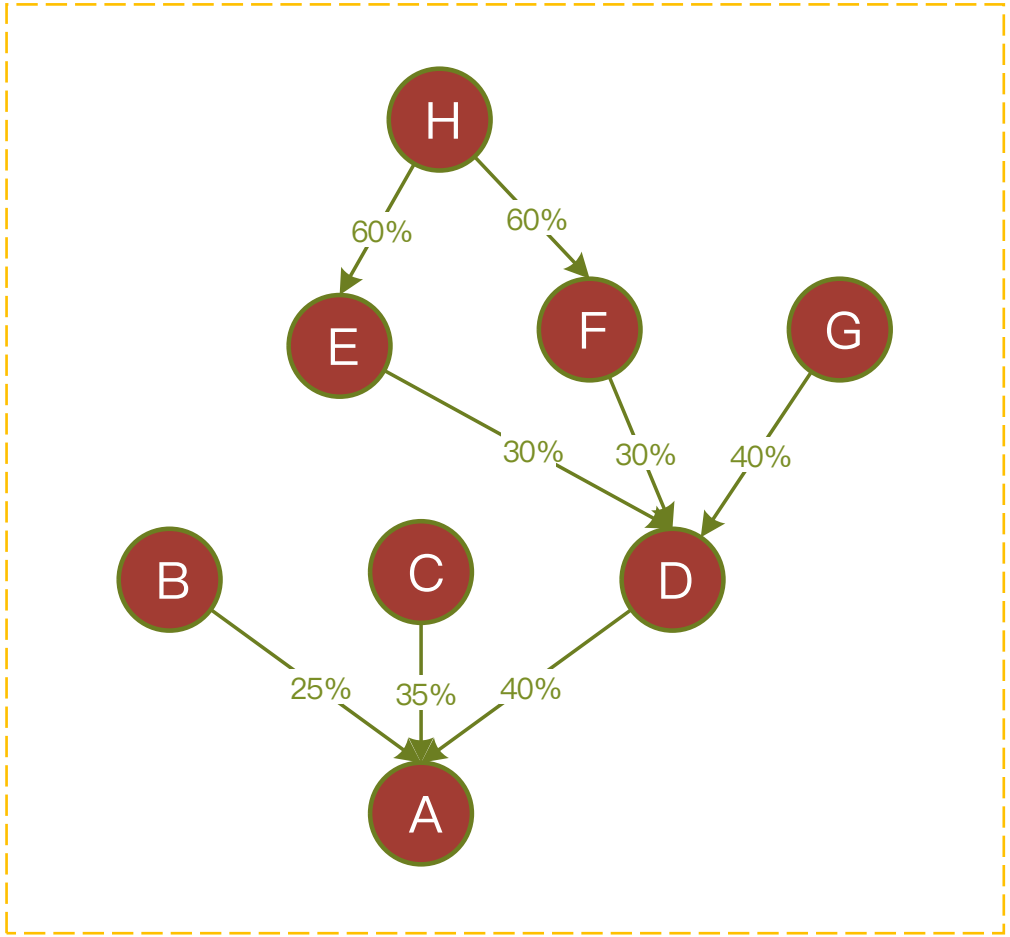
Standard Group Constructed-Suspected Real Controller Algorithm

Company

Equity paths

The suspected real controllers

A



G

Direct shareholding method

Disadvantages: Neglect of indirect shareholding

C

Cumulative shareholdings method

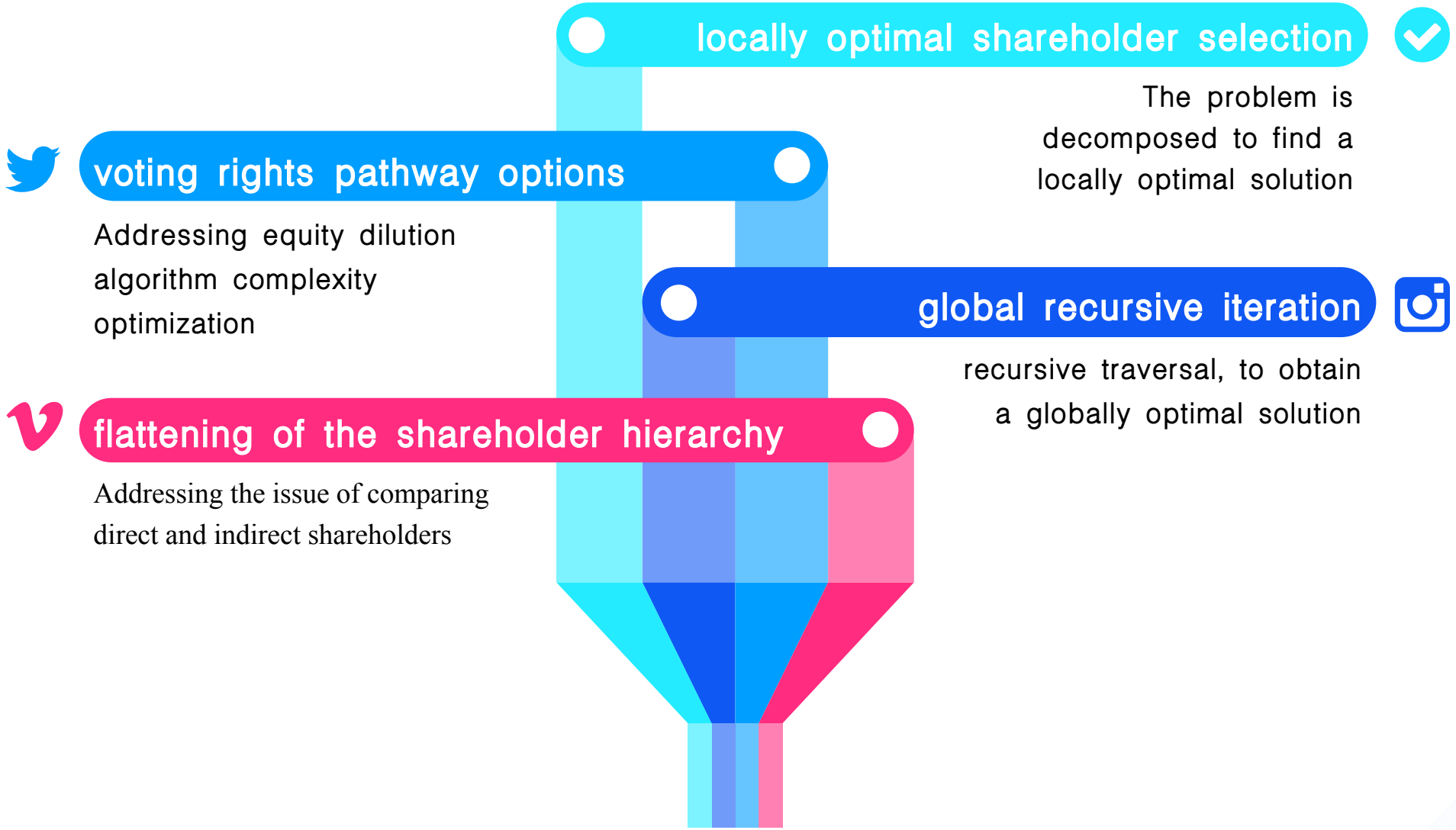
Disadvantages: Dilution by multiplication of shareholdings

H

rvoting rights iteration method

Advantage: Possesses the advantages of the direct and cumulative methods and removes the disadvantages

Standard Group Construction - voting rights iteration Method



- Core Technology 1 - Graph explore algorithms (i.e., association mapping) and computing power support: Enabling the construction of 400 million nodes and 400 million relational pairs across the entire market; and achieving unlimited hierarchical penetration of beneficial controllers.
- Core Technology 2 - Front-end visual presentation of large-scale nodes: Supporting sub-second rendering and display of over 10,000 nodes in the browser, thereby enhancing user experience.

- 300 million + nodes
- 400 million + relationships
- 1.85million of Capital groups

- the largest membership size of the group faction : 9,084
- the deepest levels : 26

证书号第5959523号

发明专利证书

发 明 名 称：风险关联企业的聚类方法、装置、设备及存储介质

发 明 人：黄佳;胡新;郑虎;童灵馨;马园园;陈锐

专 利 号：ZL 2022 1 1098554.0

专利申请日：2022年09月09日

专 利 权 人：中证数智科技（深圳）有限公司

地 址：518000 广东省深圳市前海深港合作区前湾一路1号A栋201室(入驻深圳市前海商务秘书有限公司)

授权公告日：2023年05月12日 授权公告号：CN 115170271 B

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，颁发发明专利证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。专利权期限为二十年，自申请日起算。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长
申长雨

2023年05月12日

第1页(共2页)

其他事项参见续页

证书号第5959523号

专利权人应当依照专利法及其实施细则规定缴纳年费。本专利的年费应当在每年09月09日前缴纳。未按照规定缴纳年费的，专利权自应当缴纳年费期满之日起终止。

申请日时本专利记载的申请人、发明人信息如下：

申请人：

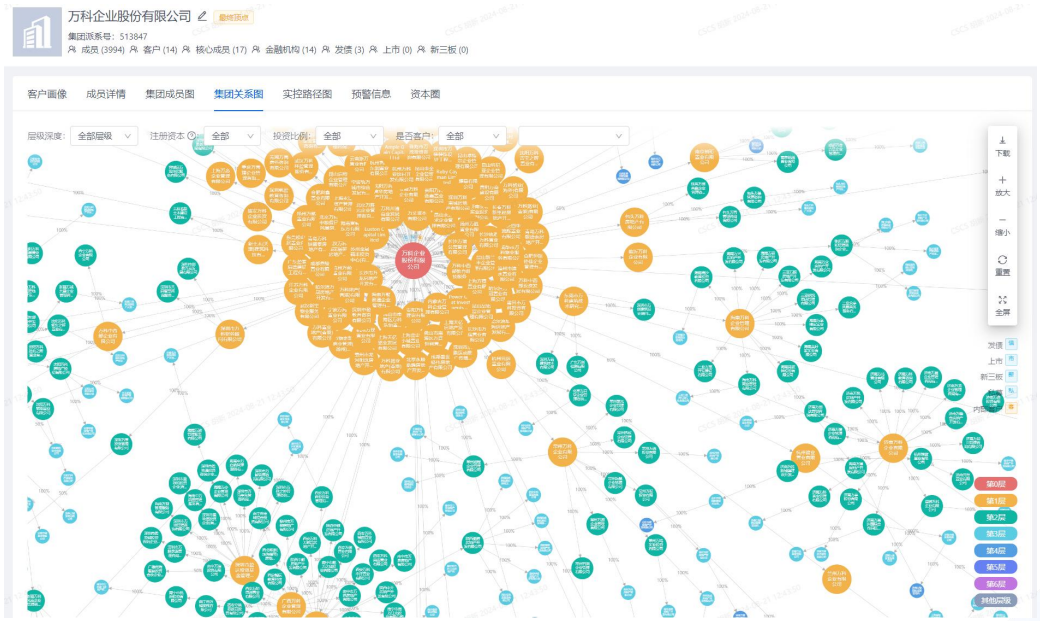
中证数智科技（深圳）有限公司

发明人：

黄佳;胡新;郑虎;童灵馨;马园园;陈锐

第2页(共2页)

- 100,000 member size group computing time: about 2 minutes

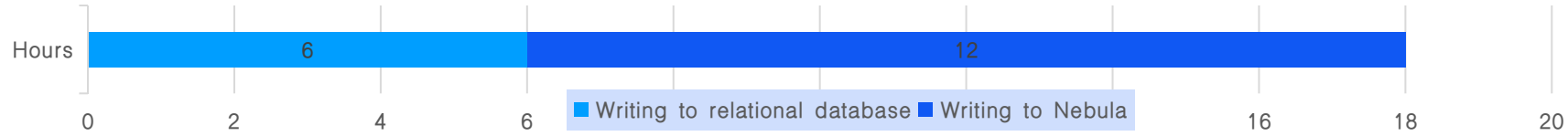


Advantages of Captical Group Using Repetitive Penetration Method

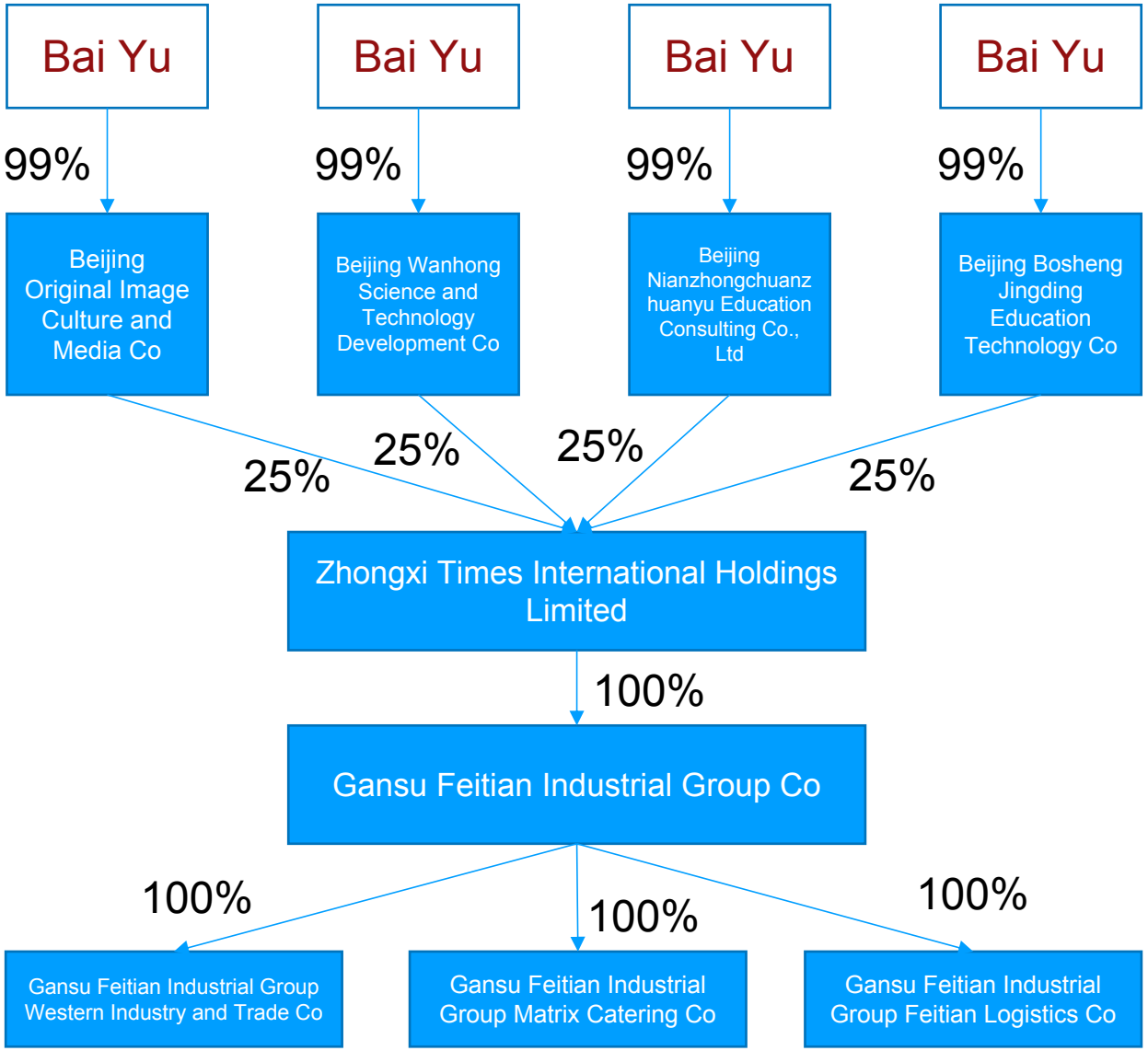
Apply technical means to meet the requirements of risk management practice for accuracy, coverage and efficienc

Value Points	Traditional Solutions	Innovative Project Solutions
Accuracy	1. Business algorithms: The direct shareholding method ignores the cumulative shareholding of indirect shareholders, leading to errors; the cumulative shareholding method makes it impossible to compare the shareholding ratios of direct and indirect shareholders at different levels due to equity dilution caused by too long shareholding paths. 2. Penetration levels: Penetrating limited levels (mostly within 10 levels). 3. Calculation scale: Considering the penetration and analysis of key Clients and subjects, limiting the number of nodes, and unable to cover the entire market.	1. Business algorithms: A voting rights layer-by-layer iterative algorithm based on equity. By comparing the size of voting rights through layer-by-layer iteration, it penetrates, calculates and identifies the actual controller of the enterprise. 2. Penetration levels: There is no limit on the number of levels when penetrating group factions, realizing penetration calculation with unlimited levels. The deepest level of the actual results of group factions exceeds 50 levels. 3. Calculation scale: It can calculate full amount of data, with more than 10,000 faction members.
Coverage	Covering the penetration and analysis of key Clients and subjects: There are certain limitations in the risk discovery and management of potential Clients and non-key Clients.	Covering more than 400 million nodes of all industrial and commercial enterprises and their directors, supervisors and senior executives, forming more than 100,000 group factions.
Efficiency	Data landing + full penetration calculation (over 400 million nodes in the entire market): 2 weeks; Real-time penetration calculation: Limited by the number of levels (within 10 levels) and calculation rules, only simple calculations can be done.	Data landing + full penetration calculation (over 400 million nodes in the entire market): Within 1 day; Real-time penetration calculation: Depending on the complexity of group relations, the calculation results are generally returned in seconds, and for particularly complex groups, the calculation results are returned in minutes.

- **Market-wide data computation time**



Adjustment of Natural person for Suspected Real-Controller Calculation



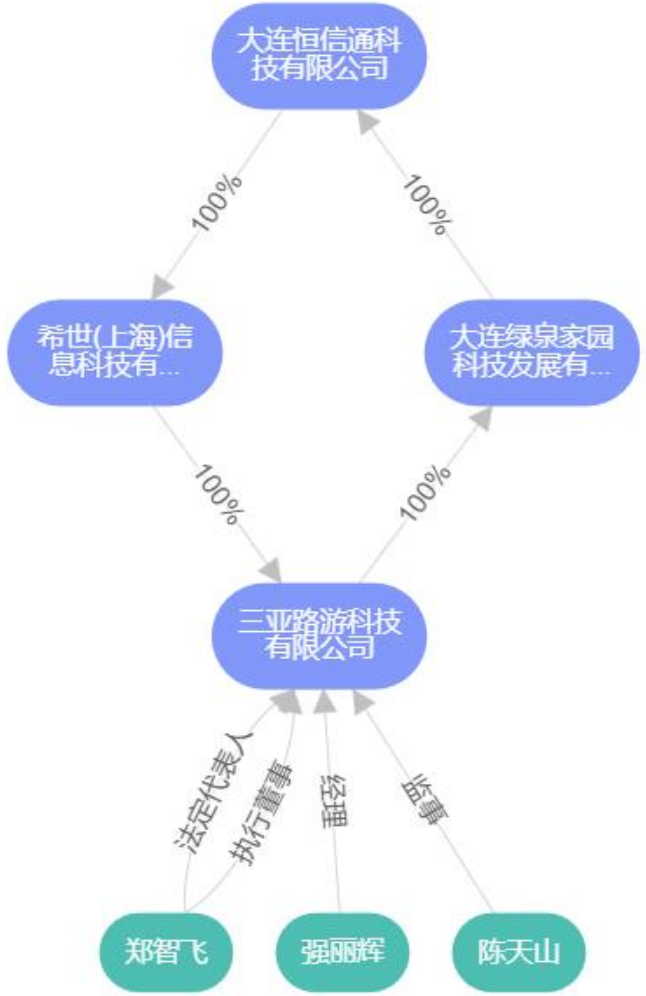
Before adjustments

It was not possible to determine whether Bai Yu, the controlling shareholder of the four companies, was the same person.

Adjusted

Considering that the top names of the suspected real controllers of the four enterprises are the same, and calculating the probability of overlap based on the structure of the graph, a natural person merger is carried out: the suspected real controller of the group is Bai Yu.

Standard Group - suspected real-control adjustments: cross-shareholding scenarios



Before adjustments

The suspected beneficial owner of Sanya Road Tour Technology Co

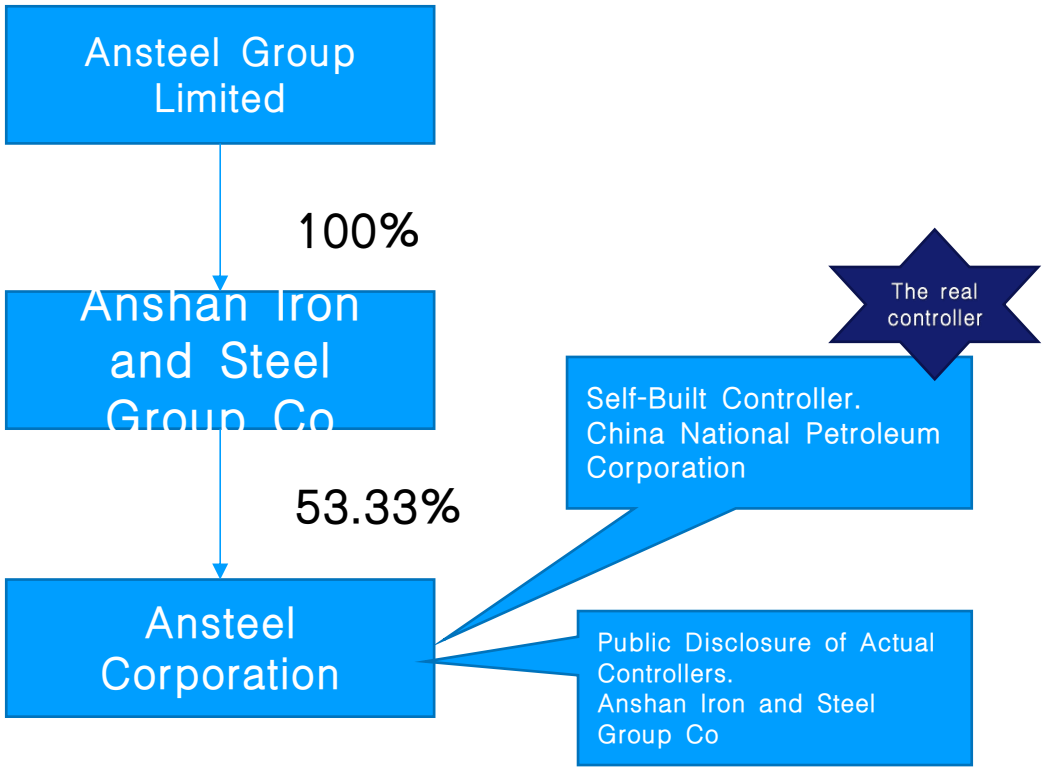
Adjusted

After considering the circular shareholding, the suspected beneficial owners are selected among them according to the following priority.

1) The node with the largest shareholding
2) the node with the largest registered capital
3) the node whose name is sorted first

The largest registered capital of the Heath (Shanghai) Information Technology Co., Ltd. is the suspected real controller.

Adjustment of top-controller for Suspected Real-Controller Calculation



- 3 sources of data on real controllers

Self-created real controller: manually modify the vertex real controller

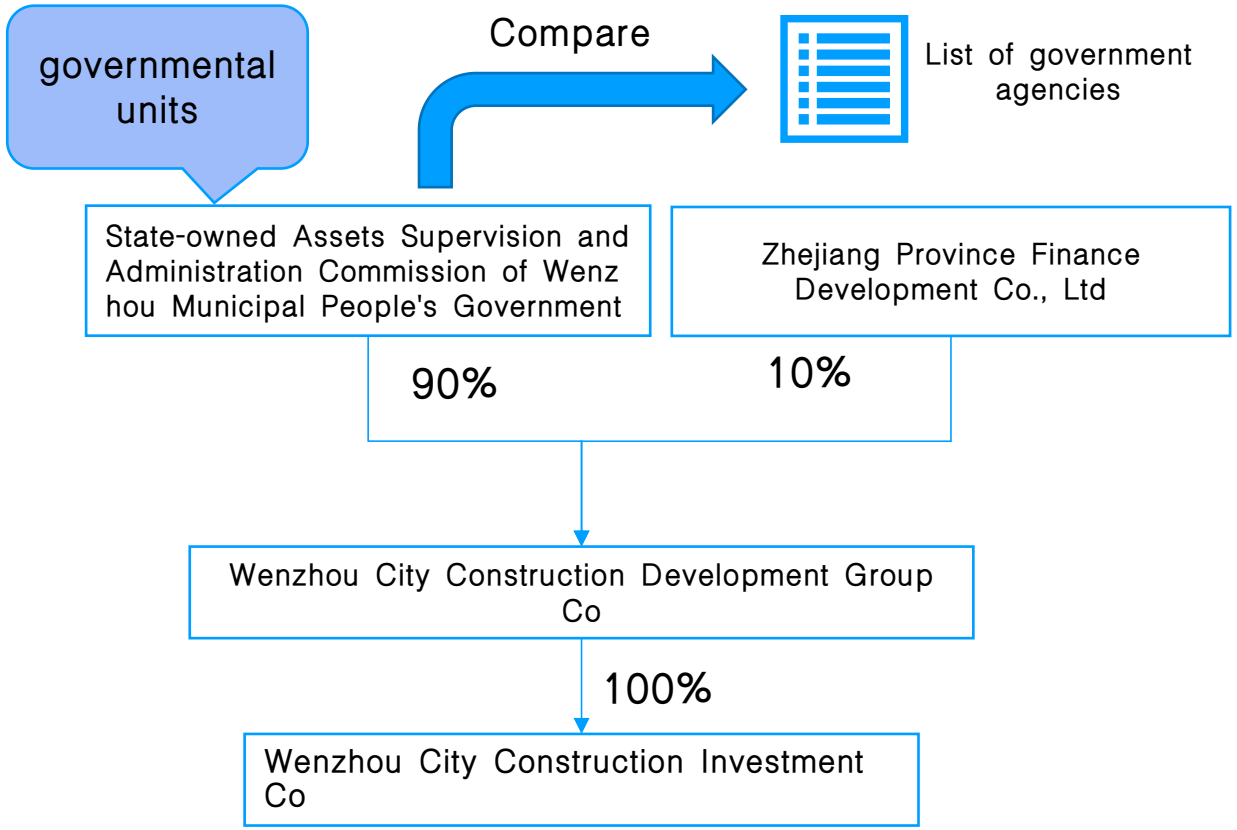
Disclosure of beneficial owners:The enterprise discloses the beneficial owners

Suspected real controller:Equity ratio of the largest apex

- Vertex Priority

self-built real controllers > publicized real controllers > suspected real controllers

Adjustment of governments for Suspected Real-Controller Calculation



The real controller is a member of the list of governmental entities, the apex back of the selection back to the first large non-governmental list of corporate shareholders.

Before adjustments

The largest cumulative shareholding of the general enterprise Wenzhou City Construction Investment Co. Ltd. based on equity penetration is the State-owned Assets Supervision and Administration Commission of the Wenzhou Municipal People's Government (90%).

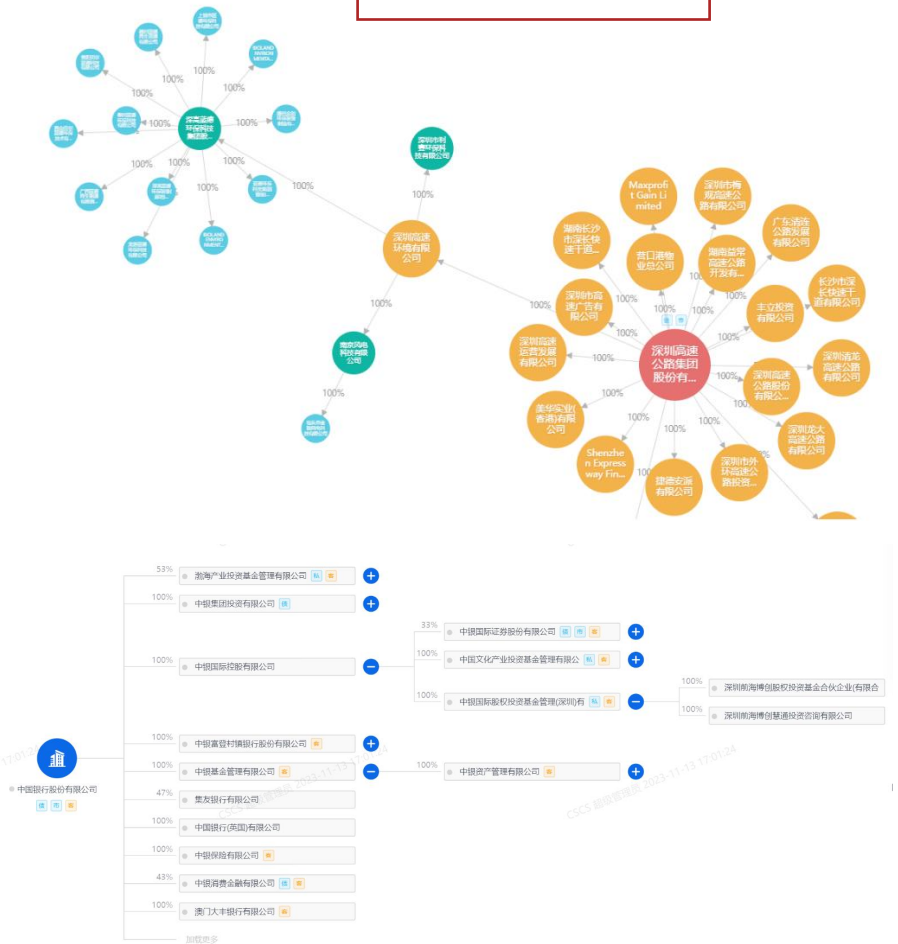
Adjusted

Considering the list of government units after the retreat, **Wenzhou City Construction Development Group Co** is the real controller (100%).

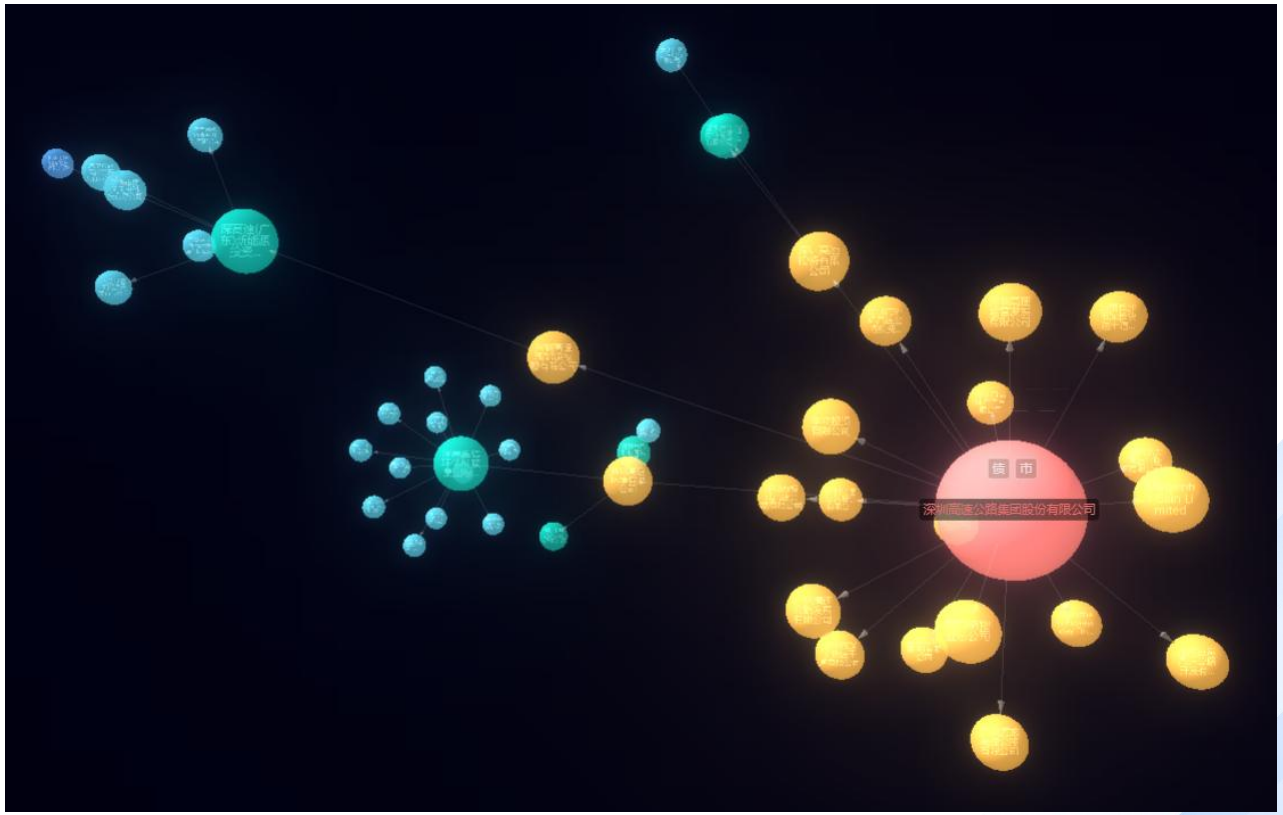
Capital Group Tree Graph

- Results for family groups based on shareholding relationships, showing members of the group factions of Shenzhen Expressway Group Co., Ltd. and Bank of China

2D Real Control Path



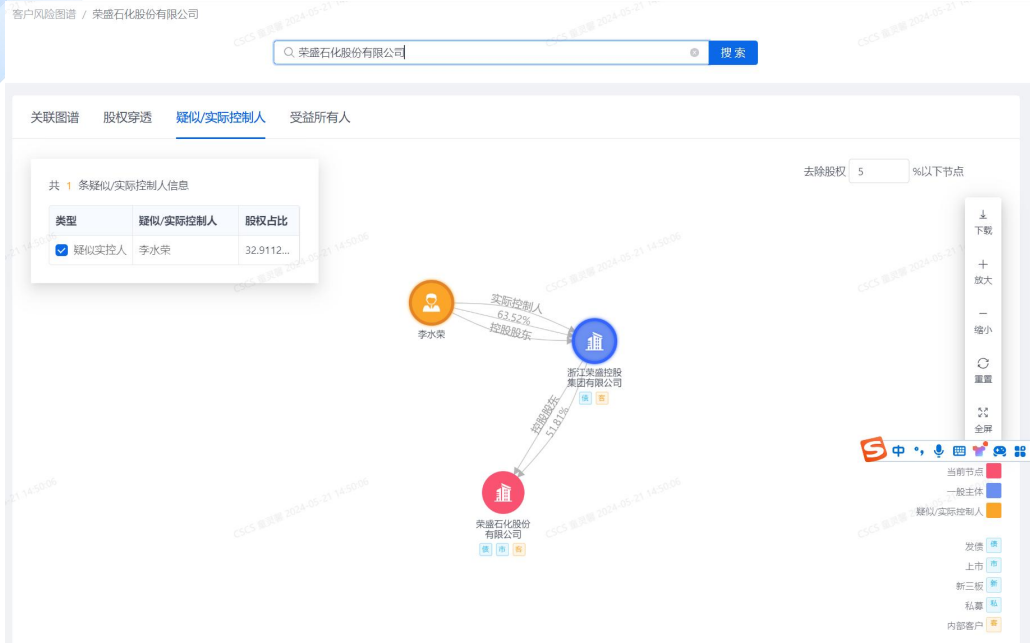
3D Real Control Path

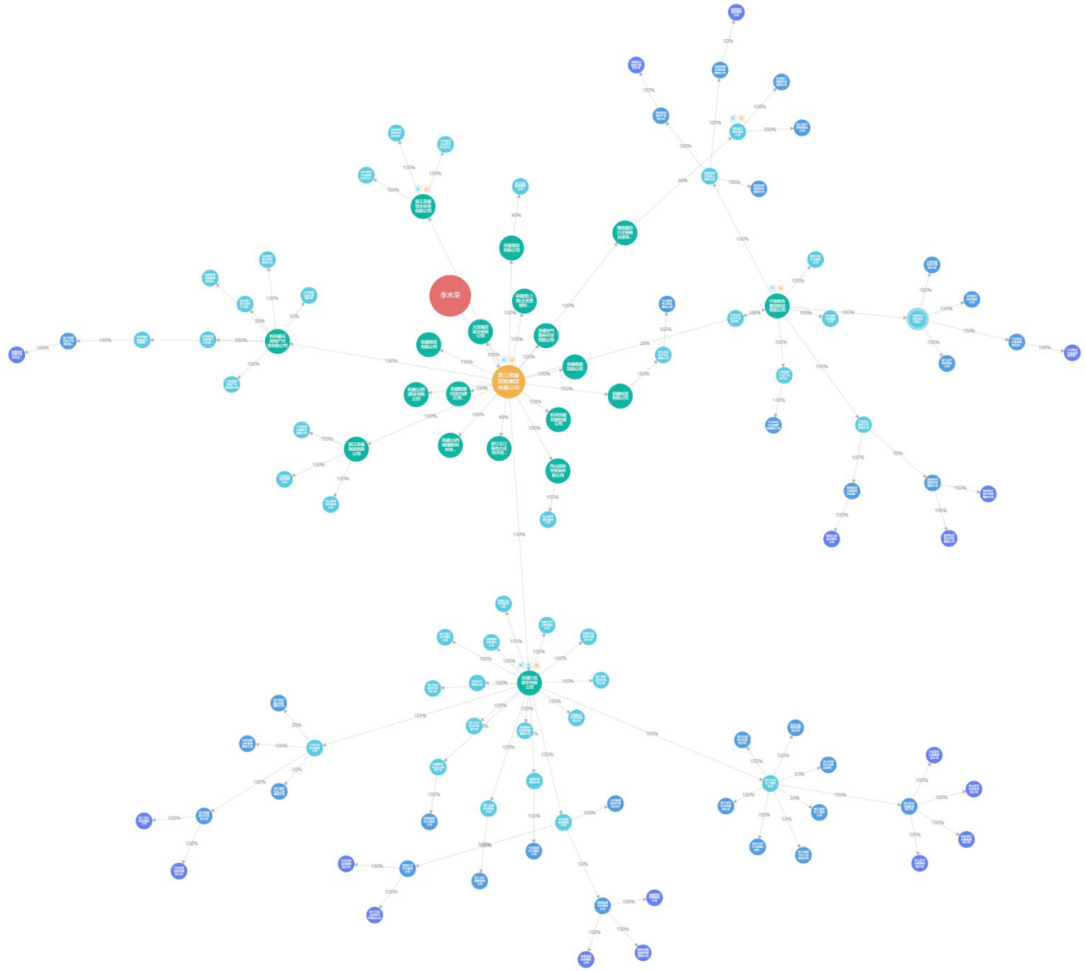


NebulaGraph

中证数智 China Securities CreditTech Services

Capital Group Tree Graph





Capital Group Tree Graph

Capital group Application Scenarios

NebulaGraph

中证数智 China Securities CreditTech Services

CONTENTS

01 About us

02 Knowledge Graph Construction

03 Graph Algorithms——Capital Group Identification

04 Knowledge Graph Application Scenarios

NebulaGraph

中证数智 China Securities CreditTech Services

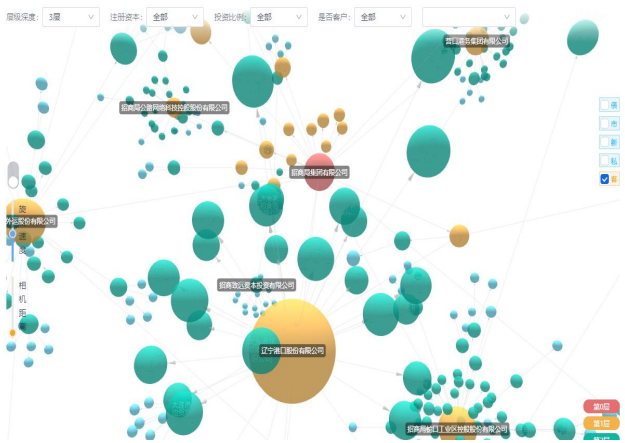
Typical Applications

Company Searching

Data from the National Enterprise Credit Information Publicity System, which includes over 120 million social entities and more than 60 million operating legal-person enterprises, allowing for precise/fuzzy queries based on fields such as enterprise name, region, and industry.

360 ° Enterprise Profile
Enterprise portrait combining external fundamentals with internal business data.

Relationship Exploring



A privatized knowledge graph capturing interconnections such as outbound investments, key customers, major suppliers, guarantees, management teams, and fund products. Dimensions are fully customizable and support multi-level drill-down to trace every relationship to its source.

Privatized Knowledge Graph
300 Million External Relationships intergrated with Internal Relationships

Risk Screening

Judicial and compliance data including case filings, court judgments, dishonesty records, hearing schedules, enforcement actions, seizures and freezes, together with integrity disclosures from regulators, ministries, and official government portals.

Accurate Adverse Information
Coverage various of risk type

Early Warning Monitoring

7×24 global media & social-listening covering 11 major source categories, with targeted monitoring of specific industries, entities, keywords, or outlets. AI-driven sentiment classification and event detection are validated by human review.

Expert rules + NLP
integration for precise negative-news tagging.

NebulaGraph

中证数智

Enterprise Big data & Knowledge Graph Application Platform

Enterprise Data Platform

Decision Assistant

Application Layer

Enterprise Profile Dashboard
- Basic Information
- Relationship Networks
- Financial Operations
- Sentiment & Legal Risk
- Business Intelligence
- Risk Indicators

Relationship Graph
- Enterprise Graph
- Equity Penetration
- Group Graph
- Controller Graph

Relationship Explore
- Realtime Explore
- Task Configuration
- Exploration Records
- Bulk Expolre

Relationship Management
- Manual Relationship
- Bulk Import
- Define Relationship
- Rule Configuration

Client Group Management
- Client Group Identify
- My Watchlists

Global Search
- Keyword Search
- Advanced Filtering

System Modules

- Metric Engine
- Rule Engine
- Process Engine
- Dynamic Page
- Smart Table
- scheduling engine
- User Permissions
- Audit Log Management

Graph Algorithm

Graph Construction
- NODE/Relationship/Property
- Pathfinding/Centrality /Community Detection

Graph-Powered Solutions
- Equity Structure Penetration
- Acting-in-concert Parties
- Ultimate Beneficiary
- Suspected Controller
- Cross-Shareholding
- Capital Group

Relationship Graph

Database

External Data
- Corporate Registry
- News & Announcements
- Media Sentiment
- Judicial Integrity Profile

Internal Data
- Business Information
- Relationship Networks
- Client Information
- Manual Data

Risk Early Warning and Investigation for Risk Management

Early Warning Management Process

1. Risk investigation
2. Risk identification
3. Risk assessment
4. Emergency handling
5. Resolution and disposal

Early Warning Signal Management

Risk Discovery

- System monitoring
- Manual reporting
- Regular inspection

Risk Identification

- Threshold classification
- Risk stratification
- Subject classification

Early Warning Management Principles

01 Early Identification

Establish an effective risk signal identification and risk information discovery mechanism to intervene in risky Clients early.

02 Rapid Response

Deal with risk early warning signals in a timely manner, including verifying and judging their authenticity, impact, reporting in a timely manner, and initiating emergency plans for high - risk signals.

03 Effective Evaluation

Effectively and accurately evalution the level of new early warning subjects and dynamically adjust the level of existing early warning subjects through on - site and off - site investigations.

04 Continuous Tracking

Continuously track and report the risk status of early warning subjects until the risks are resolved.

Using Graph Algorithm to Risk Transmission Management

Scenario 1

Regular investigation of relationships between risk subjects and Clients

Input 1：Risk enterprise group：All defaulting entity in the market

Input 2：My Client group/Portfolio

Input 3： Default association relationship configuration (can be reset)

Output：There are 9 subjects in the Client group that are related to risk subjects, which may have potential risks and need to be investigated.

Other Scenarios

- Supplier bid - rigging verification
- Existing business verification of new business subjects
- Related transaction verification

NebulaGraph

中证数智

Using Big Data & Graph for Marketing

Information Sources

Public Market Information

- News
- Announcements
- Research reports
- Industrial and commercial information
- Information leaderboard
- .....

Internal Information

- Project documents
- Weekly reports
- ...

Business Opportunities

for Listed Companies

- Share increase
- Share reduction
- Share repurchase
- Securities lending
- ...

for Private equity institutios

- Potential overseas private placements for domestic filing
- Private fund managers with excellent performance
- ……

Intelligent Recommendation

Business Tags

What kind of business the business opportunity corresponds to

Department Tags

Which departments can conduct related businesses

Client Tags

Whether it is already a company Client, and the contact information of the corresponding Client manager

Associated Client Algorithm Recommendation

Graph explore algorithm to establish a contact graph and reduce Client acquisition costs
Information Sources

Using Big Data & Graph for Marketing ----Detailed Procedure

Business Opportunity Tag Identification

- Step Description: Identify business opportunity tags from financial texts, such as stock reduction business opportunities
- Model: BERT + Classifier
- Large Model Application: Automatic labeling, expanding the corpus, enriching the test set, improving model training efficiency and classification accuracy

Business Opportunity Information Extraction

- Step Description: Identify entities (such as enterprises, securities, etc.) corresponding to business opportunity tags and relationships between entities (such as merger and acquisition relationships)
- Model: GRU + CRF for entity recognition, BERT + Classifier for relationship extraction
- Large Model Application: Automatic labeling, expanding the corpus, enriching the test set, improving model training efficiency and classification accuracy

Graph Construction

- Step Description: Store business opportunity tags, entities and relationships in a knowledge graph

Business Department Preference Portrait Generation

- Step Description: The large model parses the historical documents of the business department (such as emails, meeting minutes and reports) into structured data, and analyzes to obtain the business department preference portrait
- Model: Large model

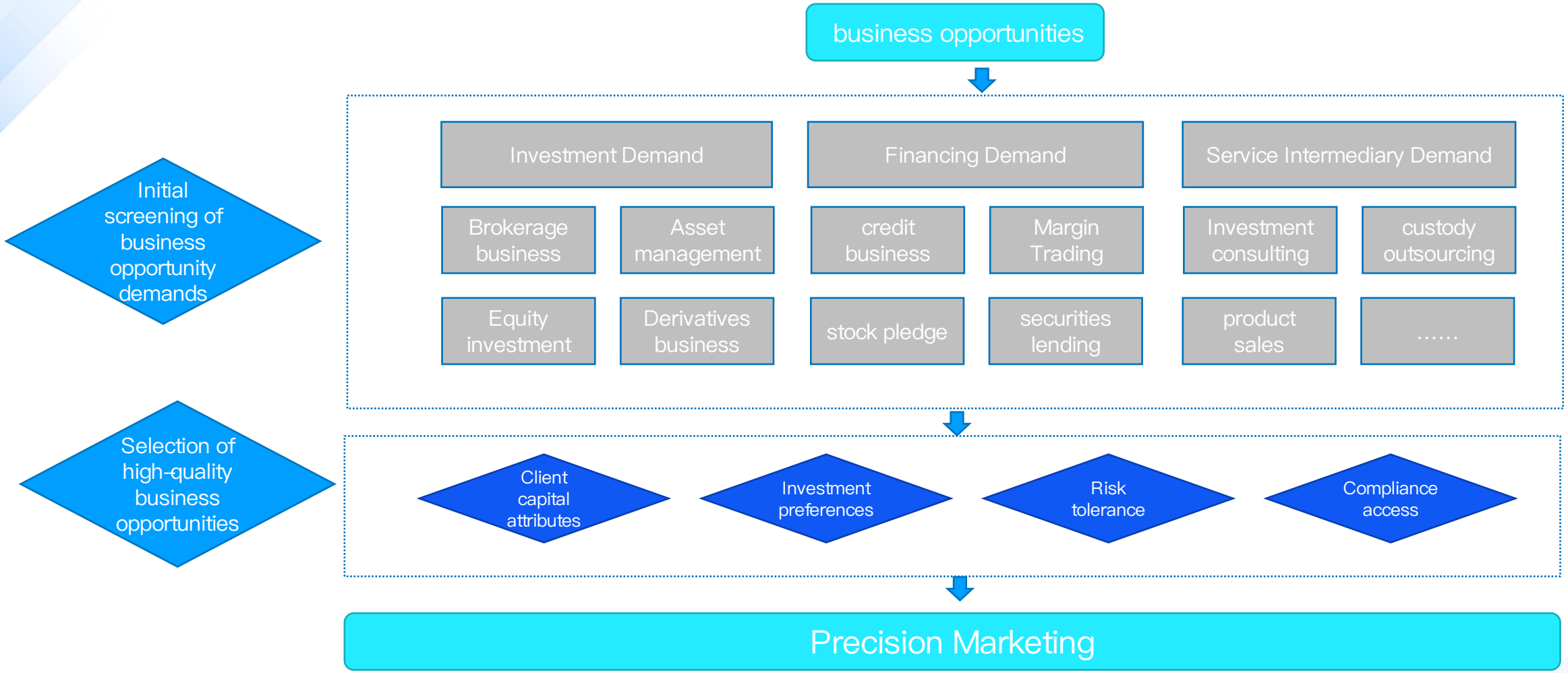
Business Opportunity Recommendation

- Step Description: The preference portrait is matched with the data in the knowledge graph to recommend personalized business opportunities for the business department. For example, if the business department prefers to invest in a certain type of enterprise, the recommendation system will find investment opportunities related to these enterprises in the knowledge graph and recommend them to the business department
- Model: Recommendation algorithms, such as collaborative filtering, content-based recommendation

Sending Recommendation Results to Business Departments

- Step Description: The generated recommendation results are sent to the business department through an automated report generation and distribution system

Precision Business Opportunity Matching



NebulaGraph

中证数智

Using Graph to Exploring Business Opportunity

NebulaGraph

https://nebula-graph.com.cn

THANKS

https://nebula-graph.com.cn

GitHub: vesoft-inc/nebula

Twitter: @NebulaGraph

Facebook: @NebulaGraph

https://discuss.nebula-graph.com.cn

本人微信号

Email:huxin@ Chinacscs.com

Https://www.chinacscs.com

Fax: 0755-83653315

Tel: 0755-84362888

Adress:深圳市南山区深湾二路82号神州数码国际创新中心东塔43楼

中证数智微信公众号

Apache Doris

The Fastest Analytics and Search Database for the AI Era

Kang Xiao

Apache Doris PMC Member & VeloDB VP

What's Apache Doris

Apache Doris: The Fastest Analytics and Search Database for AI

DORIS

2013 Project Creation

2017 Open Source

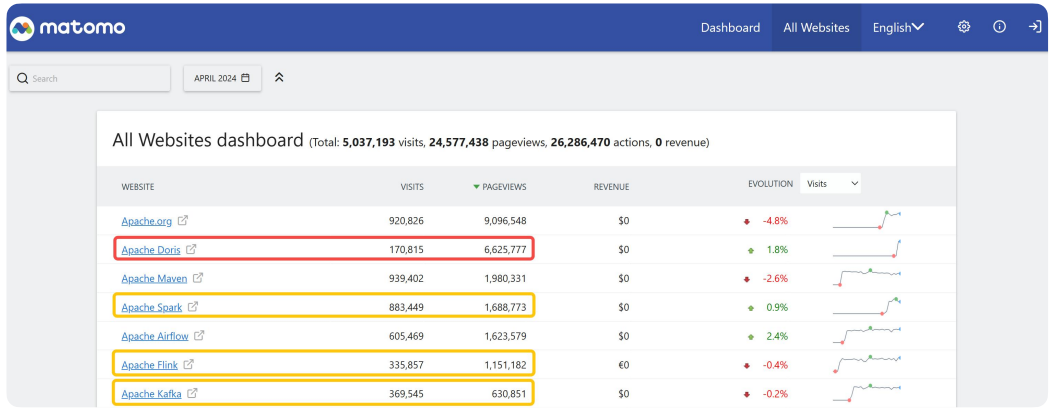
2022 ASF Top Project

13k+ GitHub Stars

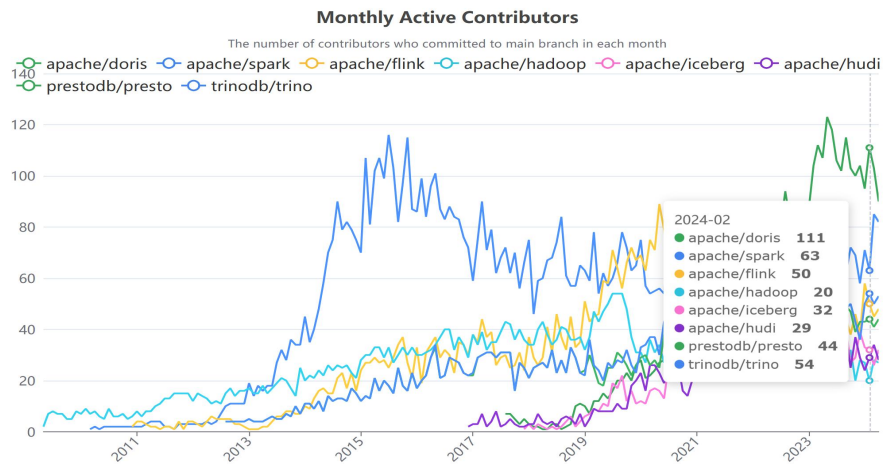
640+ Contributors

5000+ Enterprises

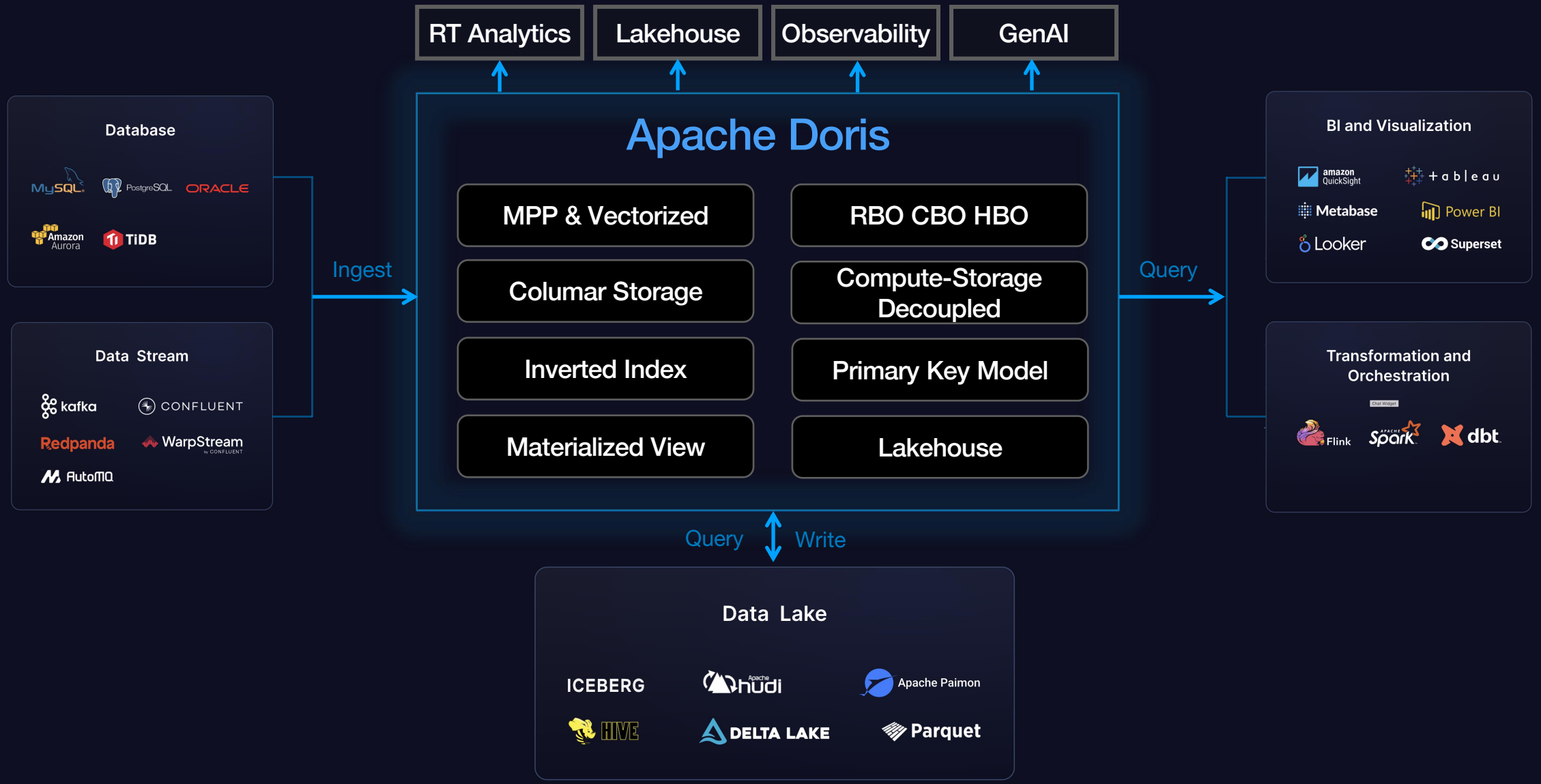
#1 Daily Page Views of Apache Project Websites



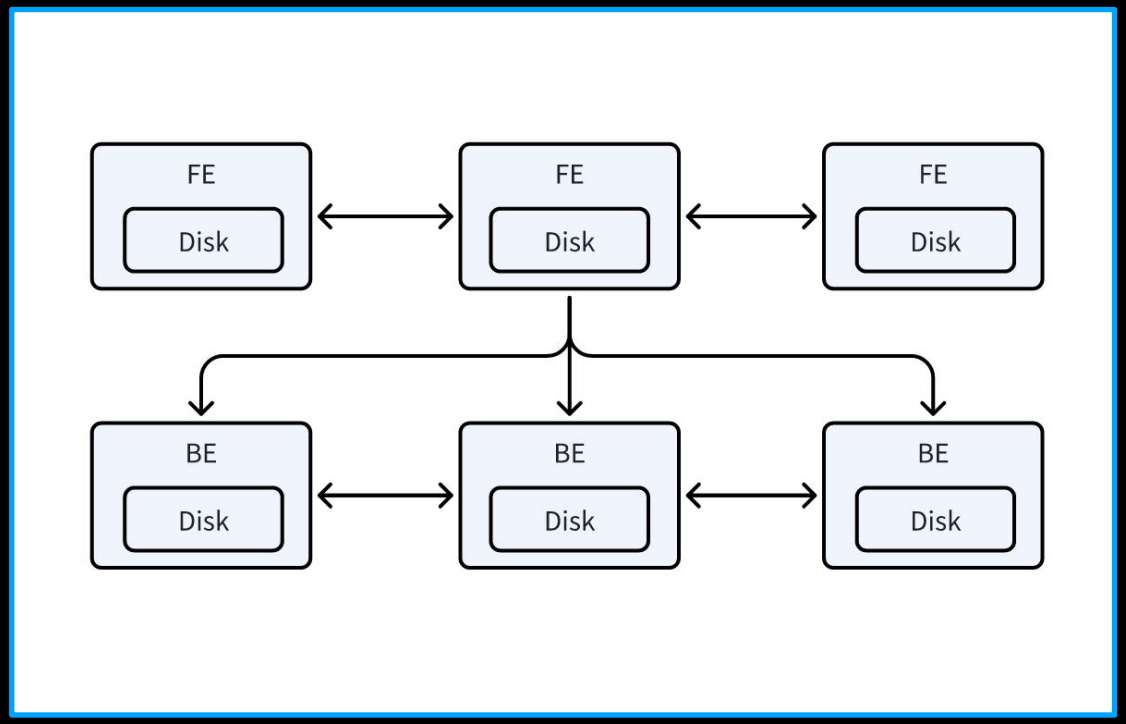
#1 Monthly Active Contributors



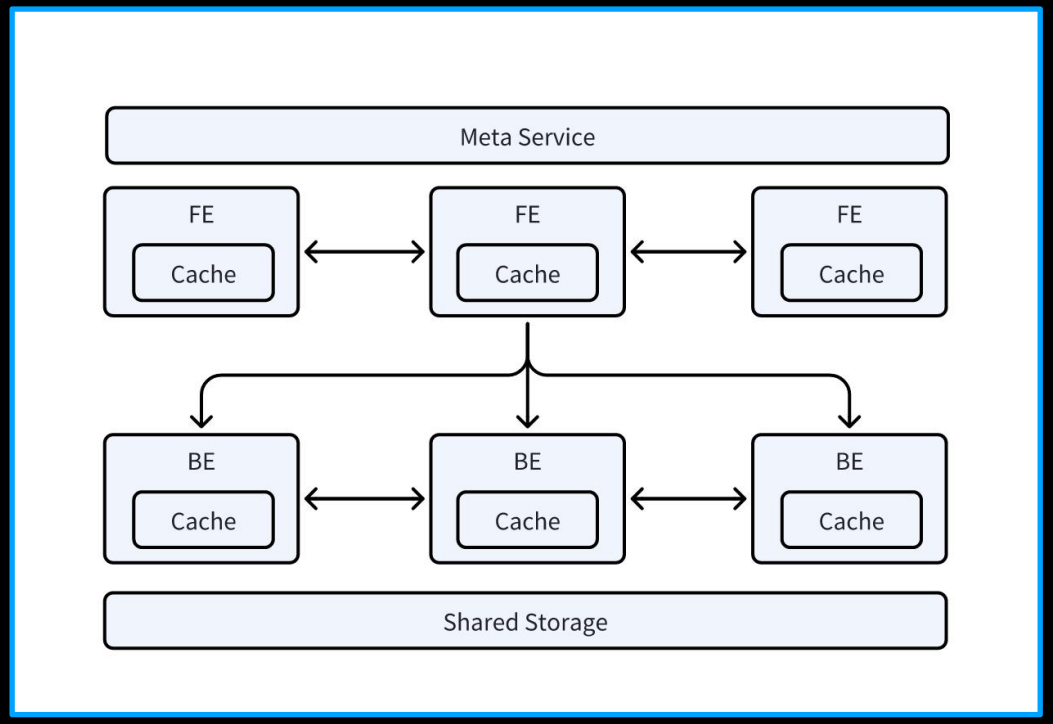
Product Archtecture



Technical Archtecture



Compute Storage Coupled Mode



Compute Storage Deoupled Mode

Simple: Only FE and BE

Reliable: no dependency

Low latency: local disk

Elastic: compute and storage scale independently

Shared Storage: one copy of data with low cost

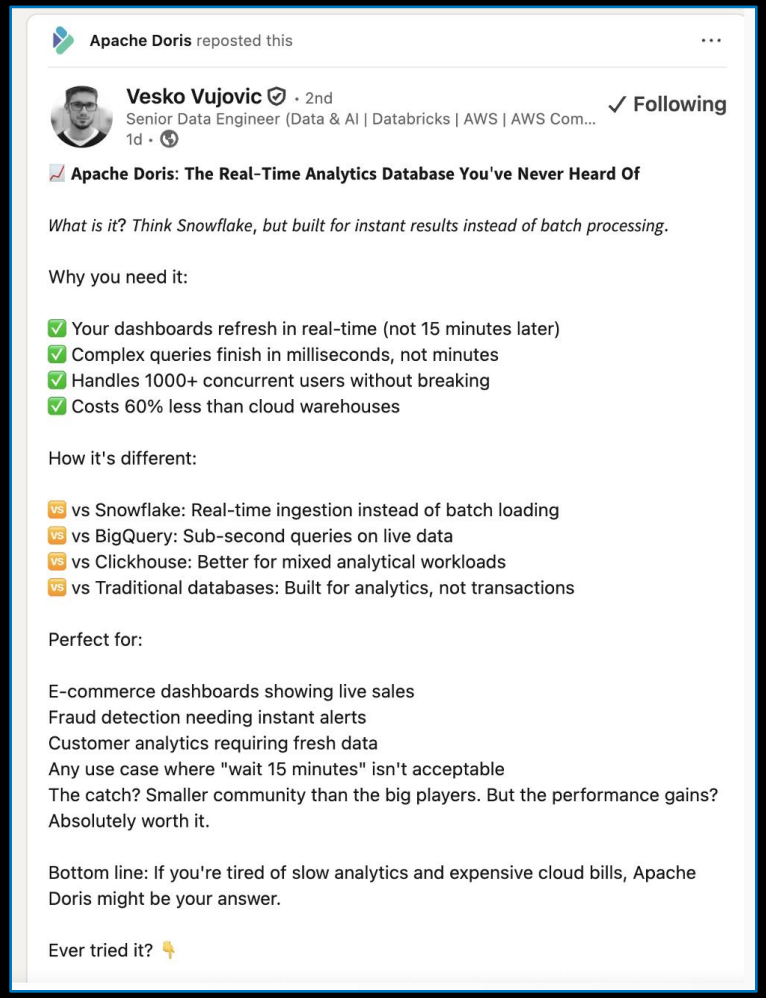
Workload Isolation: business / read write isolation

Trusted by Global Fortune 500

Over 5,000 enterprises worldwide adopt Apache Doris in their core data analysis scenarios

Why Apache Doris

Why Doris



Apache Doris reposted this

Vesko Vujovic · 2nd
Senior Data Engineer (Data & AI | Databricks | AWS | AWS Com...
1d ·

✓ Following

📈 **Apache Doris: The Real-Time Analytics Database You've Never Heard Of**

What is it? Think Snowflake, but built for instant results instead of batch processing.

Why you need it:

✅ Your dashboards refresh in real-time (not 15 minutes later)
✅ Complex queries finish in milliseconds, not minutes
✅ Handles 1000+ concurrent users without breaking
✅ Costs 60% less than cloud warehouses

How it's different:

🆚 vs Snowflake: Real-time ingestion instead of batch loading
🆚 vs BigQuery: Sub-second queries on live data
🆚 vs Clickhouse: Better for mixed analytical workloads
🆚 vs Traditional databases: Built for analytics, not transactions

Perfect for:

E-commerce dashboards showing live sales
Fraud detection needing instant alerts
Customer analytics requiring fresh data
Any use case where "wait 15 minutes" isn't acceptable
The catch? Smaller community than the big players. But the performance gains? Absolutely worth it.

Bottom line: If you're tired of slow analytics and expensive cloud bills, Apache Doris might be your answer.

Ever tried it? 👇

Why you need it

- Your dashboards refresh in real-time (not 15 minutes later)
- Complex queries finish in milliseconds, not minutes
- Handles 1000+ concurrent users without breaking
- Costs 60% less than cloud warehouses

How it's different

- **vs Snowflake**: Real-time ingestion instead of batch loading
- **vs BigQuery:** Sub-second queries on live data
- **vs ClickHouse:** Better for mixed analytical workloads
- **vs Traditional databases:** Built for analytics, not transactions

Perfect for

- E-commerce dashboards showing live sales
- Fraud detection needing instant alerts
- Customer analytics requiring fresh data
- Any use case where "wait 15 minutes" isn't acceptable

Why Doris

Fastest

agg, join and search

Cost effective

storage and compute

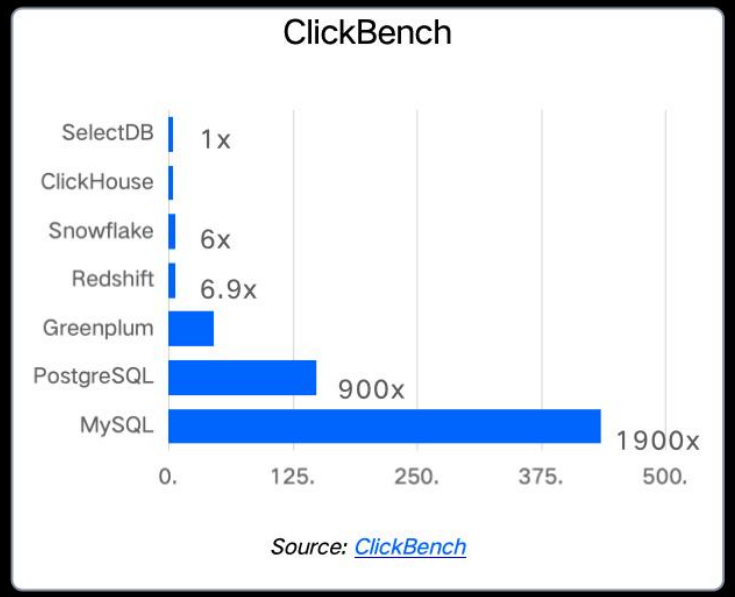
Unified

one engine for analytics

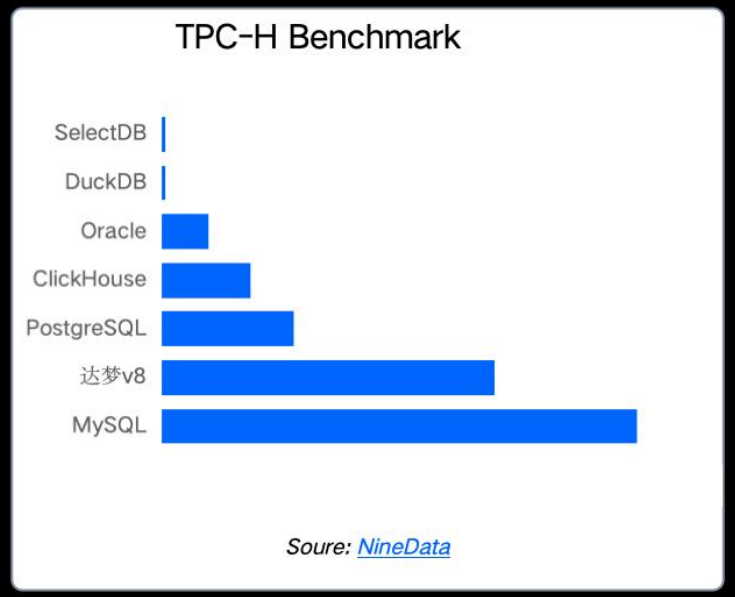
Run anywhere

bare metal, vm, k8s, cloud

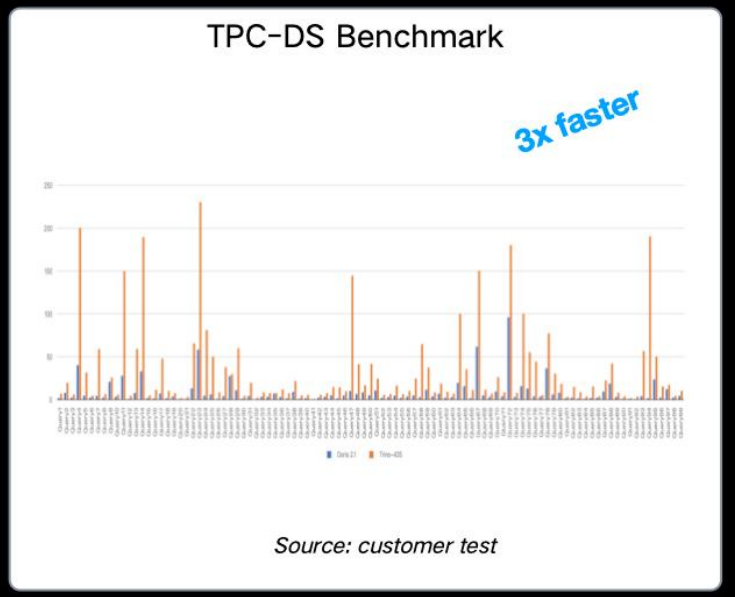
Fastest



Source: ClickBench



Soure: NineData



Source: customer test

Vectorized

- SIMD instructions，both x86 and ARM
- less virtual function call and cache miss

CBO Optimizer

- cost based join reorder，pushdown，RF
- high concurrent query optimizations

Rich Index

- skip index：bloom filter, min/ max/ sum
- point index：prefix index，inverted index

Materialized View

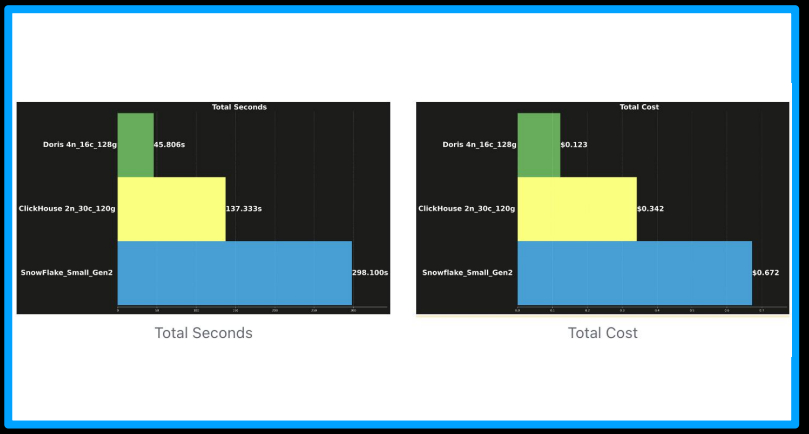
- consistent single table agg mv
- multi-table general mv

Cost Effective

CoffeBench 1.4B

5x to Snowflake

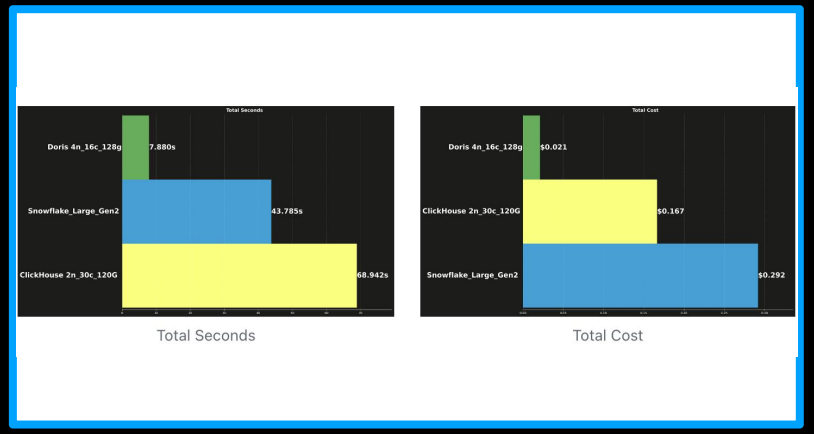
3x to ClickHouse



TPC-H-SF100

50x to Snowflake

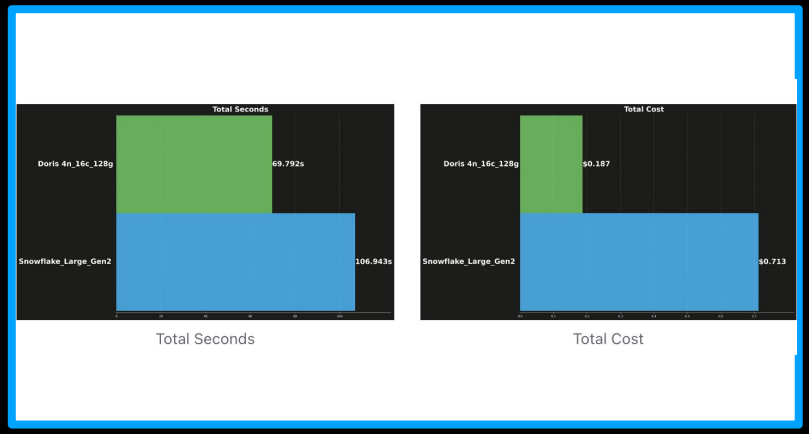
8x to ClickHouse



TPC-DS-SF100

4x to Snowflake

ClickHouse can not run



Unified

Real-time Analytics

faster than Clickhouse, esp JOIN

Lakehouse

3x faster than Presto / Trino

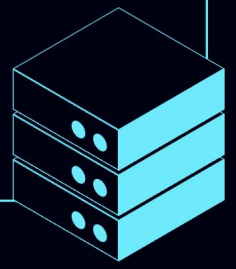
Search

10x cost effective than Elasticsearch

Deploy Your Way, Run anywhere

Apache Doris / VeloDB Enterprise

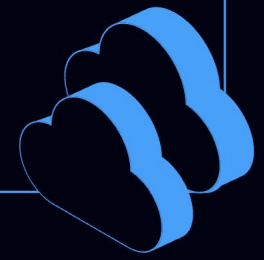
Self-managed software on premises, on VMs, or K8s



VeloDB Cloud

Fully managed, cloud-native, real-time data warehouse service

阿里云 腾讯云 亚马逊云科技 aws Azure Google Cloud



Top Use Cases of Apache Doris

Top Use Cases

Real-time Analytics

Lakehouse

Observability

AI

Real-Time Analytics

Real-Time OLAP Database

~ 1 s	< 100 ms	> 10,000 QPS
minimum data latency	average query latency	maximum query concurrency

Real-Time Ingestion & Update

- Streaming ingestion from Database CDC and Kafka
- Row-level updates with strict primary key consistency
- Data Warehousing and BI

Blazing-Fast Analytics

- MPP distributed architecture
- Vectorized execution engine
- Complex join queries with a CBO optimizer
- Runtime Filtering

High-Concurrent Queries

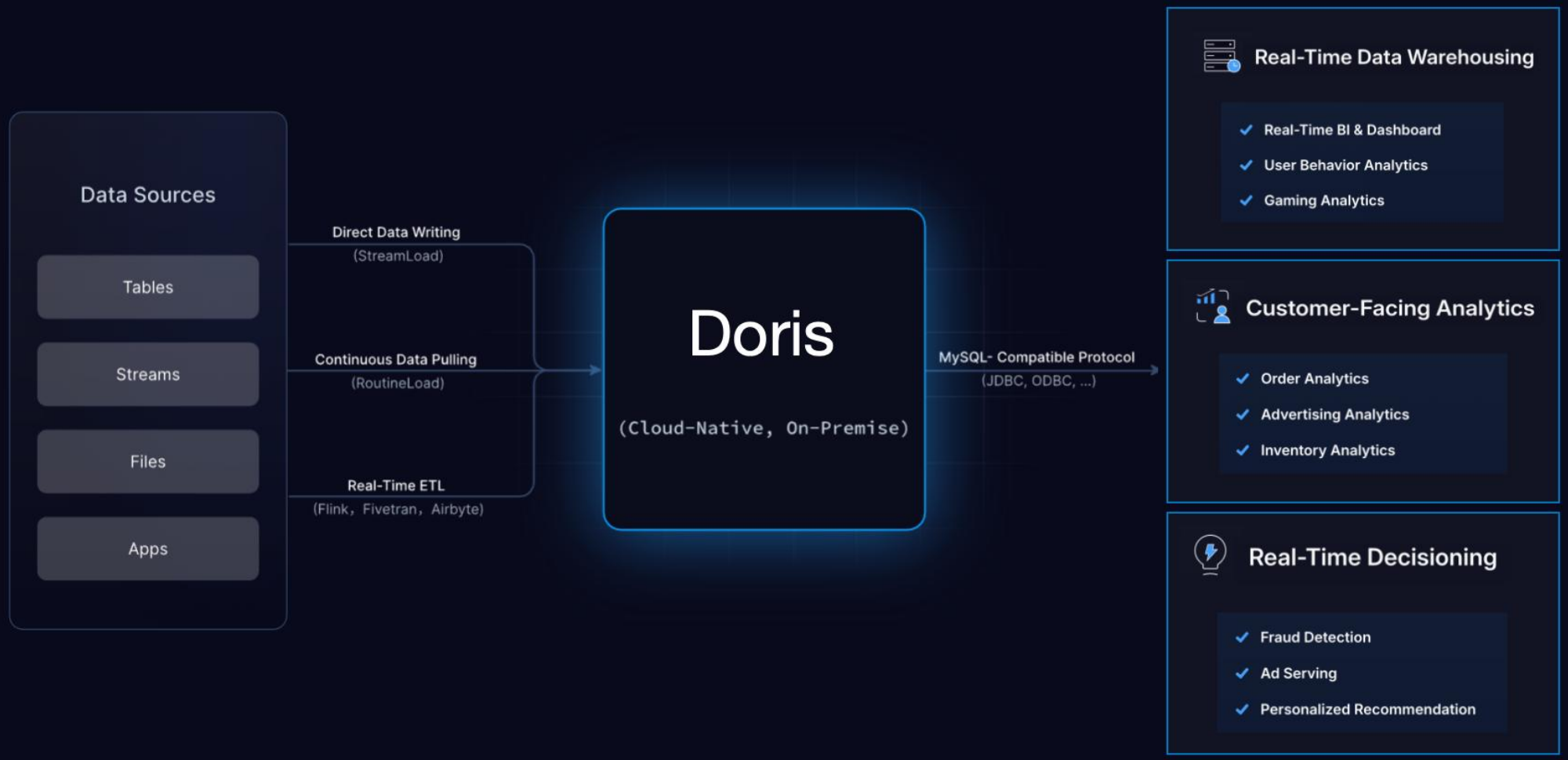
- Data pruning based on partitioning and bucketing
- Various data indexes
- Pre-aggregated table
- Materialized views

Alternative to ClickHouse

- #1 Complex JOIN Performance
- Consistent real-time ingestion and updates
- Higher concurrency performance

Real-Time Analytics

Real-Time OLAP Database



Real-Time Analytics

Real-Time OLAP Database

"Used in cell performance traceability scenarios, this system enables tracing via barcodes. It involves joining up to ten tables, with large tables containing 10 billion rows. By leveraging inverted indexes and other optimizations, query performance is accelerated by more than 10x."

BYD

"Compared to the original OLAP database, query performance has improved 5-10 times, concurrency has doubled, and analysis time has dropped from 10 minutes to under 1 minute for 90% of cases, all while using just one-third of the original resources."

ZTO (NYSE: ZTO)

"Apache Doris is now widely used across dozens of business units at Xiaomi, including advertising, growth analytics, data dashboards, digital finance, e-commerce, and user profiling. The largest single cluster scales to 100 nodes and PBs."

XIAOMI

Lakehouse Analytics

The Fastest Lakehouse SQL Engine

Blazing-Fast

With a high-performance query engine and fast metadata/data caching, VeloDB becomes the fastest lakehouse analytics engine, outperforming Trino by 2-3 times.

Open

Seamlessly integrates with mainstream open data formats and catalogs in the lakehouse ecosystem, while also providing extensive support for other data sources, including databases.

Unified

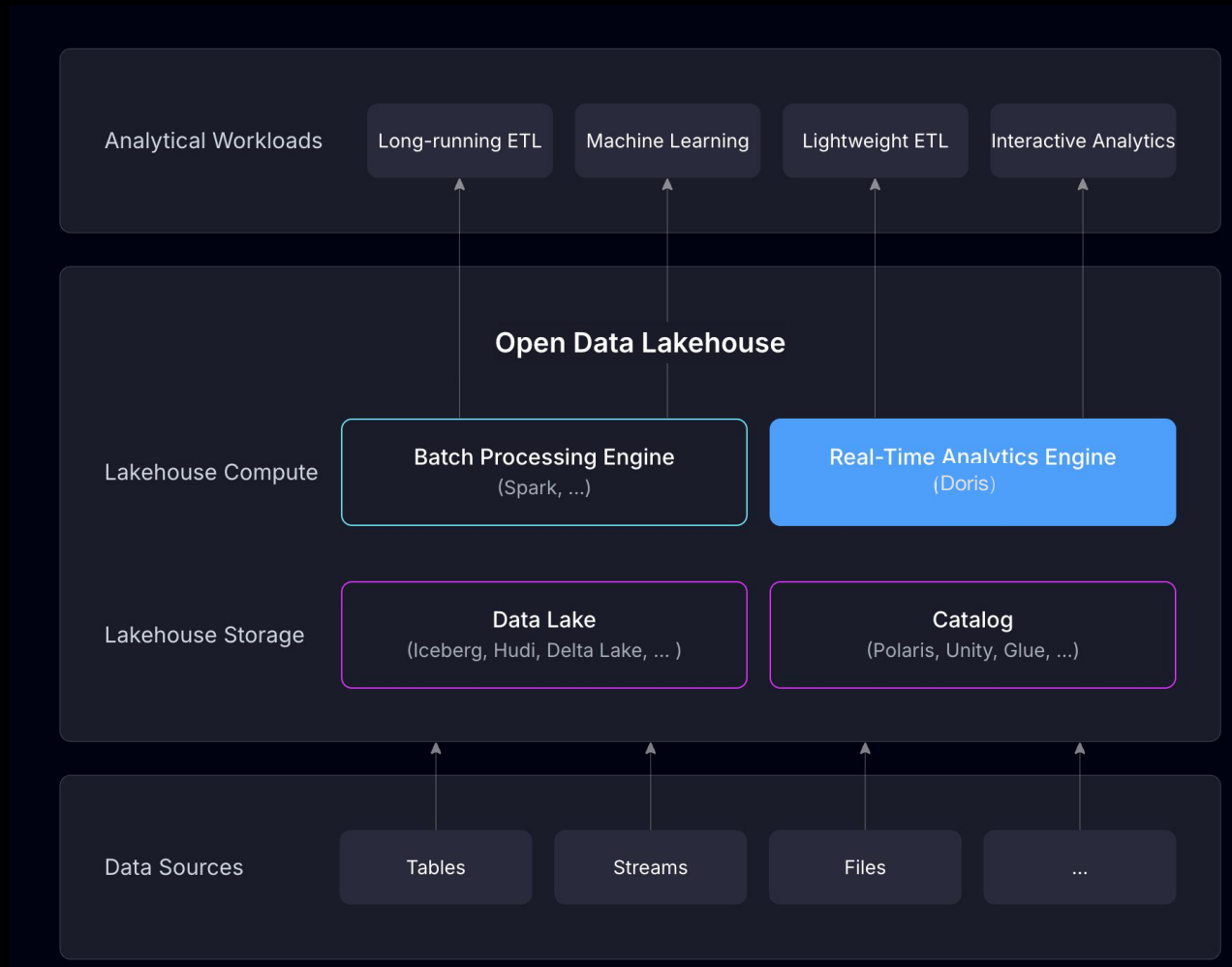
With built-in storage, VeloDB can be used as an analytical database, a lakehouse analytics engine, or both, unlocking more powerful capabilities.

Alternative to presto trino

- 3x faster than Trino / Presto as lakehouse
- 10x faster than Trino / Presto as warehouse
- Better together for Lakehouse and warehouse

Lakehouse Analytics

The Fastest Lakehouse SQL Engine



Real-Time Analytics Engine

Use Doris as the real-time analytics engine, primarily responsible for supporting interactive analytics and lightweight ETL computational workloads.

Batch Processing Engine

Use Spark-like batch processing engines, primarily responsible for supporting long-running ETL and machine learning computational workloads.

Open Lakehouse Storage

Build an open lakehouse storage based on Data Lake using open table formats and open Catalog.

Lakehouse Analytics

The Fastest Lakehouse SQL Engine

"After introduced Doris to replace Presto, with an average daily query volume of more than 1 million, the P95 performance has been improved by nearly 3x, and the computation resource savings up to 48%, which is a significant benefit."

SF EXPRESS

"We built a unified lakehouse architecture based on Apache Doris and Iceberg, enabling seamless data interoperability between Doris and Iceberg, supporting over 200 projects, with a total data volume at the PB scale and more than 15 million queries per day."

NetEase Games (NASDAQ: NTES)

"We introduced Apache Doris to replace Trino and Pinot, unified the data management on PostgreSQL, Elasticsearch and Iceberg. This significantly simplified our architecture, improved query performance and and reduced resource costs by 30%."

eSurfing Cloud

Logging and Observability

High Performance, Low Cost, Open Observability Stack

10x Cost EffectiveCompared to Elasticsearch

- Cut storage volume by 80% while keeping inverted index.
- 5x ingestion rate with inverted index
- 2x full-text search performance

Efficient and Flexible JSON Data

- VARIANT data type extract JSON fields as sub-columns without ETL
- 10:1 compression ratio, 3x compared to text or binary JSON
- 8x analytical performance

Seamless Integration with Observability Ecosystem

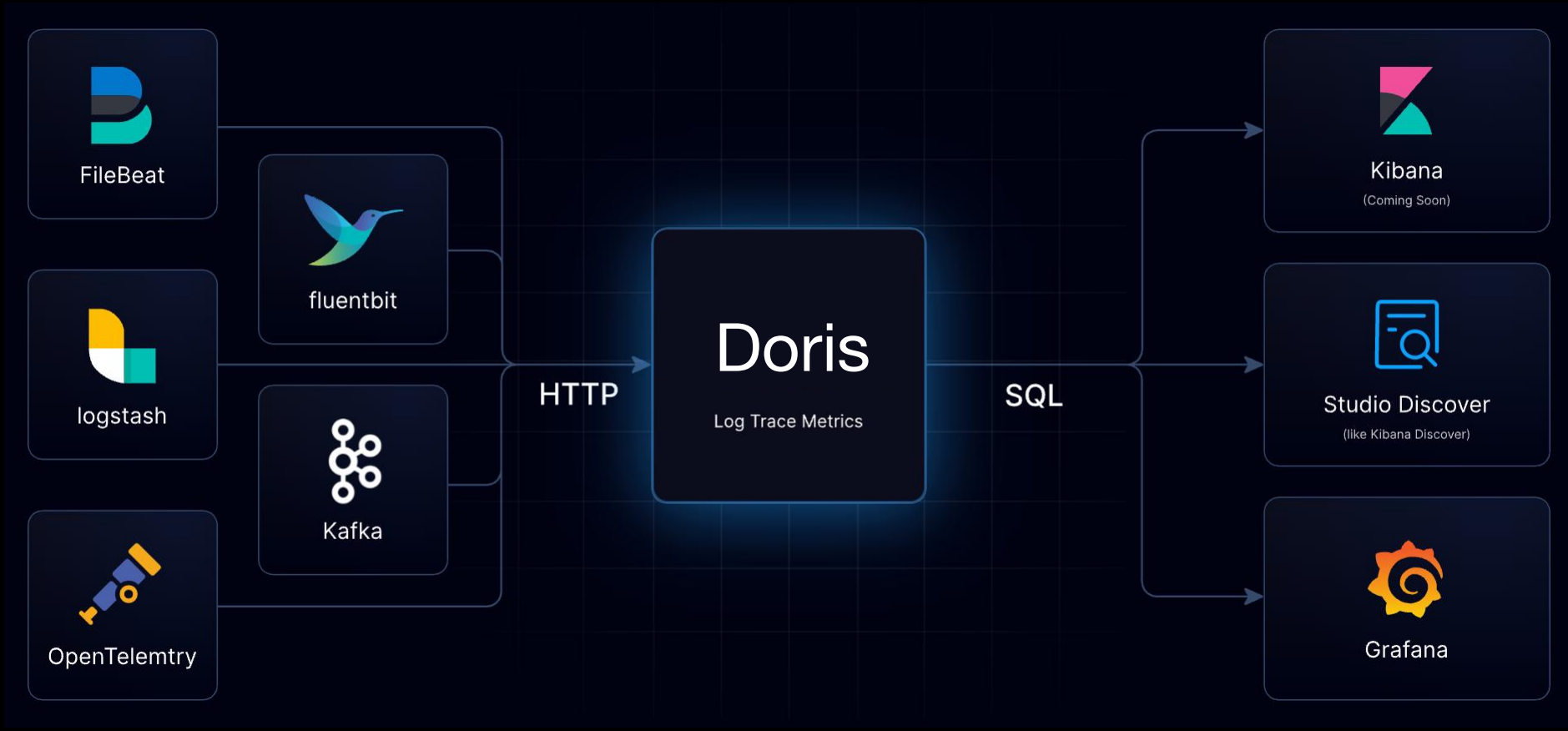
- Logstash collector and Kibana visualization in ELK
- OpenTelemetry exporter and Grafana visualization
- AI Observability with Langfuse and OpenLLMetry

Alternative to elasticsearch

- ✔ 10x Cost effective
- ✔ 5x Ingestion throughput
- ✔ Open observability ecosystem

Logging and Observability

High Performance, Low Cost, Open Observability Stack



Logging and Observability

High Performance, Low Cost, Open Observability Stack

"We generate 14 billion log entries daily, totaling 80TB, with a total archive exceeding 40PB. Our old platform based on Elasticsearch suffered from high storage costs. Switching to Apache Doris cut costs by 50% and speed up queries by 2 to 4 times."

CITIC BANK
(OTCMKTS: CHBJF)

"At Netease, high storage costs and slow queries on Elastcisearch were serious issues. Now, Apache Doris has cut storage costs by 70%, let us use SSDs for hot data at no extra cost, and sped up queries 11x with much less CPU usage."

NETEASE (NASDAQ: NTES)

"We Built a fault analysis platform based on Apache Doris. Compared to Apache Pinot and Elasticsearch, the solution reduced the number of servers from 64 to 14, decreased storage space by 80%, and improved query performance by 40%."

VeloDB for AI

The AI-Ready analytics databases

Agent-Facing Analytics

- ✓ ~1s real-time ingestion & update
- ✓ ~100ms blazing-fast analytics
- ✓ ~10k QPS high-concurrent queries

AI Lakehouse

- ✓ Data preparation
- ✓ Feature Engineering
- ✓ Quality Evaluation

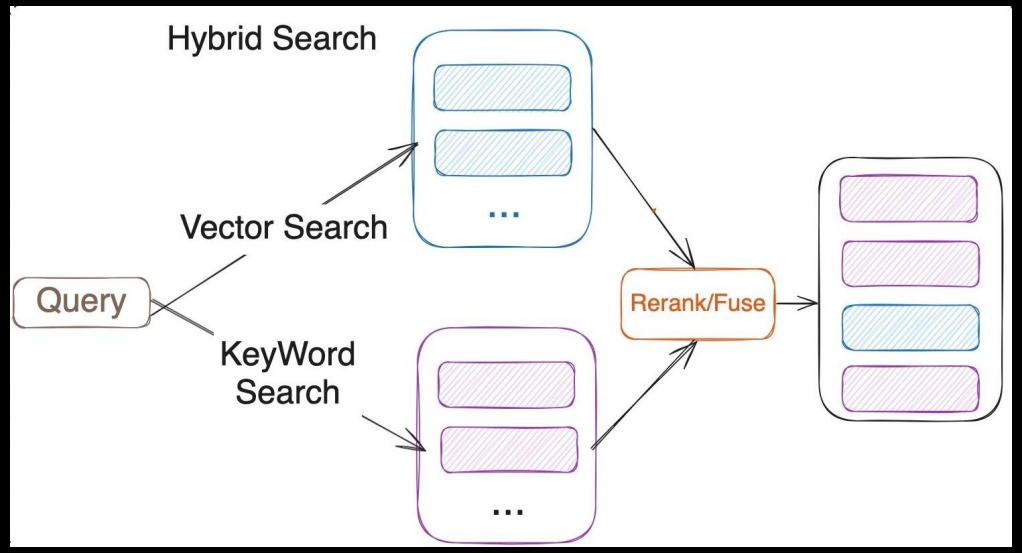
AI Observability

- ✓ data volume & cost efficiency
- ✓ high throughput ingestion
- ✓ LLM ecosystem integration

AI Unstructured Analytics

- ✓ Hybrid Search
- ✓ AI-Powered SQL
- ✓ MCP Server

Hybrid Search




```sql
SELECT
    id,
    name,
    description
FROM docs
WHERE
    update_time BETWEEN '2025-01-01' AND '2025-05-30'
    AND description MATCH 'GenAI'
    AND version >= 2.0
ORDER BY
    cosine_distance(embedding, [...])
LIMIT 100
```

Why Vector-Capable General-Purpose Databases are Better for Enterprise GenAI ?

- Lack of Hybrid Query Capabilities
- Limited Integration with Structured Data
- Operational Complexity and Increased Costs
- Large scale data volume, eg PBs

AI Operators in SQL

Integrating GenAI for Enhanced Text Analysis in Database

```sql
SELECT
    c.candidate_id, c.name,
    j.job_id, j.title
FROM candidate_profiles AS c
JOIN job_requirements AS j
WHERE LLM_FILTER(
    CONCAT('Does the following candidate match the job description?',
        'Job: ', j.jd_text, ' Candidate: ', c.self_intro));
```

```
+--------------+-------+--------+------------------+
| candidate_id | name  | job_id | title            |
+--------------+-------+--------+------------------+
|            3 | Cathy |    102 | ML Engineer      |
|            1 | Alice |    101 | Backend Engineer |
+--------------+-------+--------+------------------+
```

- LLM_SUMMARIZE
- LLM_TRANSLATE
- LLM_SENTIMENT
- LLM_CLASSIFY
- LLM_GENERATE
- LLM_EXTRACT
- LLM_FILTER
- LLM_SIMILARITY
- LLM_FIXGRAMMAR
- LLM_MASK

VeloDB for AI

The AI-Ready analytics databases

"Before, our LLM business had a slow, unstable logging system based on Loki. Now, Apache Doris manages massive logs, ensuring over 99.9% availability and sub-second query latency on 100 million logs."

MINIMAX talkie

"We use Doris as a unified OLAP engine to replace ClickHouse and Elasticsearch, combining it with large language models to provide users with a personalized, real-time, and flexible intelligent data service platform."

TENCENT MUSIC ENTERTAINMENT (NYSE: TME)

"We developed a vector index feature based on Doris and contributed it back to the open-source community. We have built a PB-scale vector storage and hybrid search service, providing RAG capabilities for numerous business lines."

ByteDance

Use Cases for Web3

A leading global cryptocurrency trading platform, established in 2018 and headquartered in Singapore, with operational centers in Dubai, Hong Kong, and other regions. The platform is committed to providing safe and efficient cryptocurrency trading services for users worldwide. It supports spot trading, derivatives trading (such as futures and options), and quantitative trading. With a daily trading volume exceeding USD 10 billion, the platform serves users across Asia-Pacific, North America, and Europe, and has surpassed 10 million registered users globally.

Business Requirements

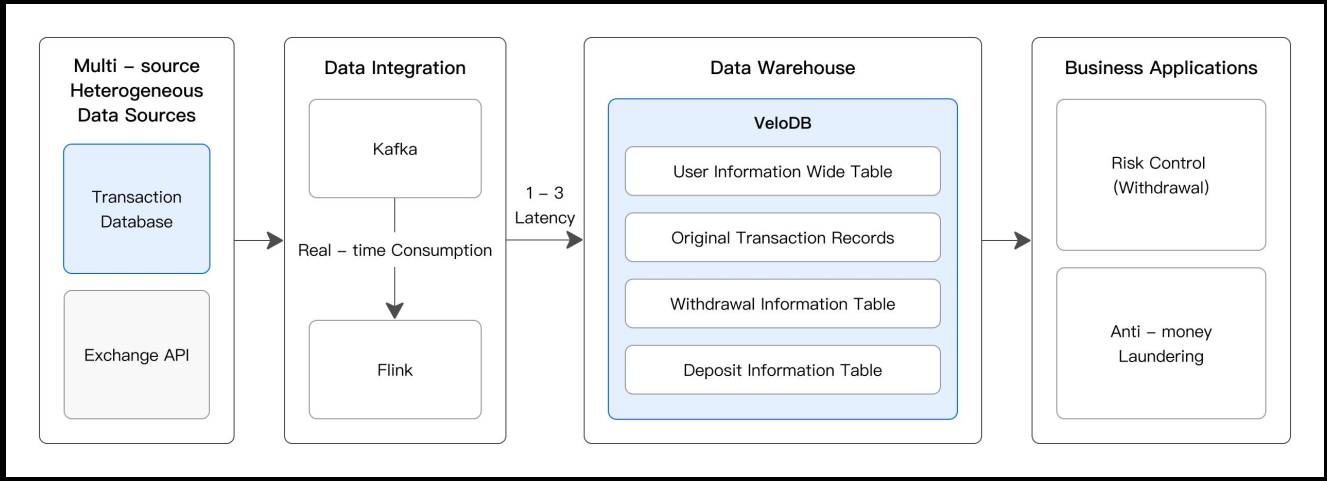
- **Risk Control for withdrawal**: Ensure fund security and reduce operational risks
- **Anti-Money Laundering**: Track user's trading behaviors for AML
- **Commission Rebates**: After registering through a referral link, part of the trading fee is returned to the referrer
- **Blockchain Transaction Tracking**: Enables precision marketing based on user activities
- **Global Trading Competition**: Large-scale trading contest
- **User Trading Behavior Analysis**: personalized analytics and report

Technical Requirements

- **Analytical Performance:** Supports sub-second query and analysis over hundreds of TB of transaction data
- **Query Concurrency:** Handles thousands of QPS per user perspective with <500ms latency
- **Data Latency:** Enables real-time ingestion of transactional data
- **High Availability:** Core business data supports high availability with AZ storage and compute
- **Architecture:** Simple and easy-to-use architecture

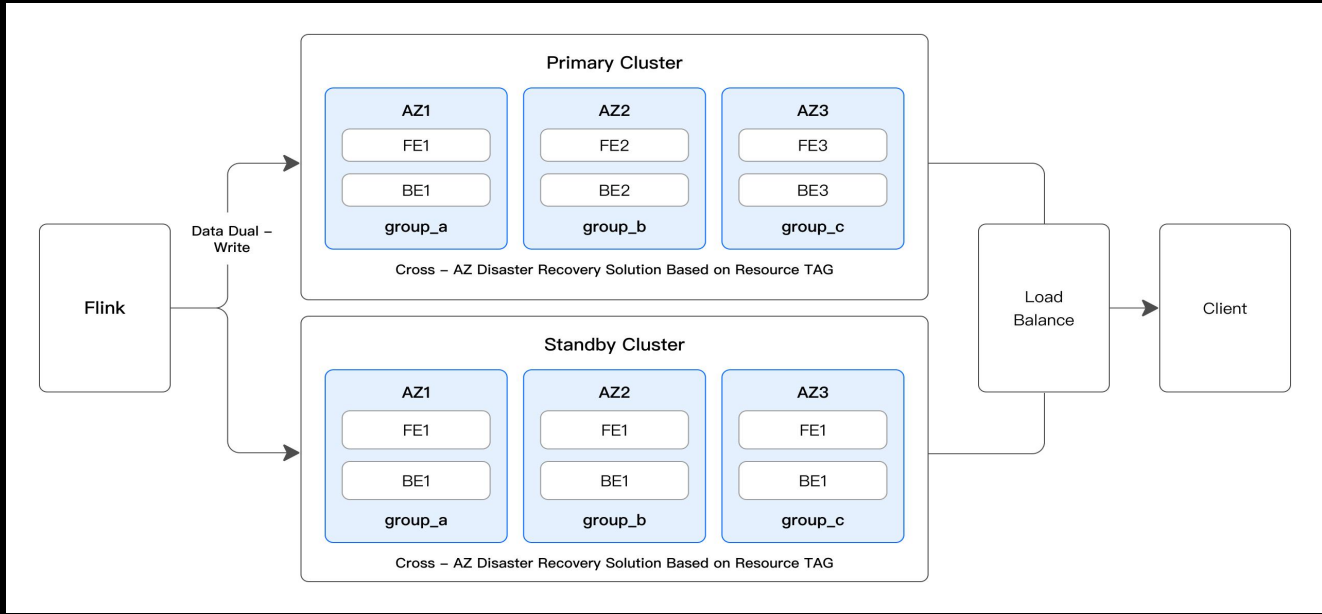
Use Cases for Web3

Scenario 1: Risk control scenario application



Application status and benefits

- **Query Type**: AD-HOC Query
- Core Business Use Cases:
  - Deposit & Withdrawal Flow Analysis
  - Anti-Money Laundering (AML) Analysis
- Query QPS:
  - Peak: 1,500 QPS
  - Stress Test: 5,000 QPS
- P95 Latency: 500 ms
- Data Volume: Several TB



Highly available deployment architecture

- Data Ingestion:

Dual-write to both primary and secondary clusters.

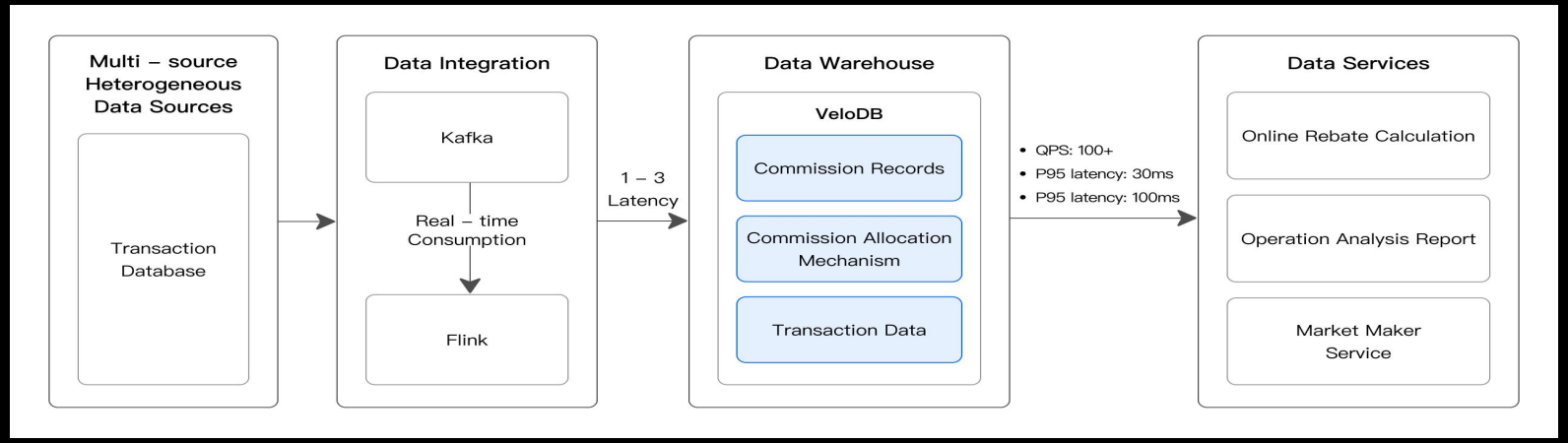
- Query Access:

Load-balanced queries across primary and secondary clusters with automatic failover.

- Cross-AZ High Availability Within Cluster:
- Frontend Nodes (FE): Deployed across multiple Availability Zones (AZs) for high availability of management services.
- Backend Nodes (BE): Deployed across AZs with Resource Tag-based replica placement to ensure each AZ maintains an independent copy of data.

Use Cases for Web3

Scenario 2: Commission Rebate Calculation



Current Status and Benefits

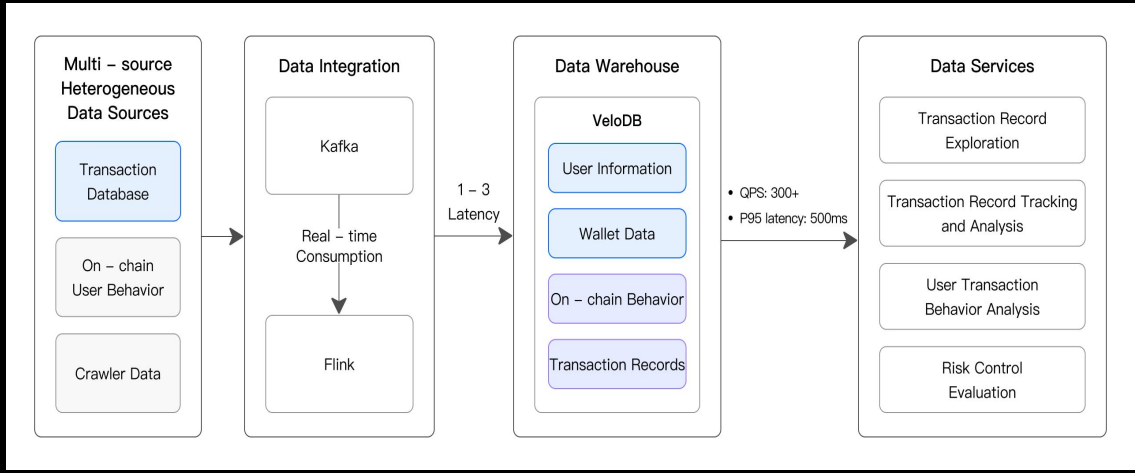
- Query Type: High-concurrency queries (real-time OLAP)
- Core Business: Rebate computation - calculating commission rebates based on user trading activity
- Query Throughput: 100+ QPS during business peaks
- Latency: P95 30ms, P99 100ms
- Data Volume: Tens of terabytes

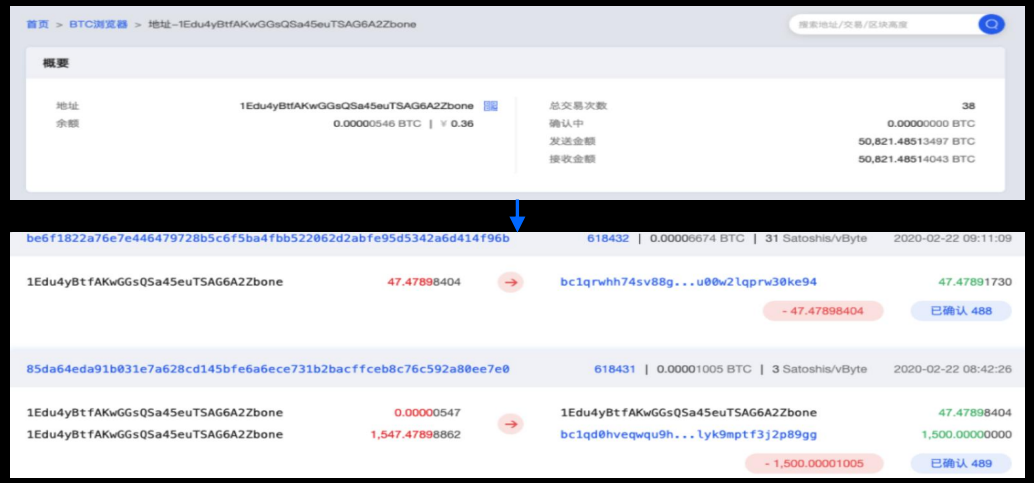
Scenarios

- Real-Time Rebate Calculation
- Operational Report Analytics
- Marketing Data Services

Use Cases for Web3

Scenario 3: Blockchain Transaction Tracking





Business Scenarios

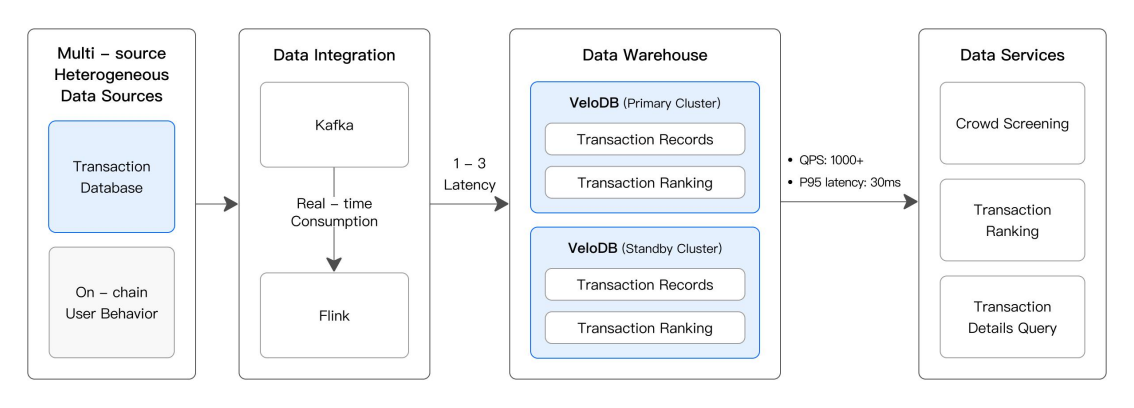
•Transaction Traceability & Forensics
Input a Transaction ID, Sender or Receiver Wallet Address to trace transaction flows on-chain. Leverage historical wallet activity and blockchain records for comprehensive transaction investigation.
•User Trading Behavior Analytics
Analyze user behavior such as trading frequency, token preferences, transaction timing, and wallet lifecycle to uncover user patterns and segment profiles.
•Risk Assessment & Scoring
Evaluate user wallets and transactions through on-chain behavioral patterns to identify suspicious activities and assign risk scores for AML, fraud detection, and compliance monitoring.

Current Status and Benefits

- Query Pattern: Ad-Hoc analytical queries
- Core Use Case: Transaction trace and behavioral analysis
- Query Throughput: 300+ QPS (Queries Per Second)
- Latency (P95): ≤ 500 ms
- Data Volume: Dozens of terabytes

Use Cases for Web3

Scenario 4: World Trading Competition



Business Scenarios

During Competition:

- •Leaderboard Ranking: Real-time ranking based on trading volume or performance.
- •Trade Detail Queries: High-frequency, low-latency queries for participants' transaction records.

Post-Competition:

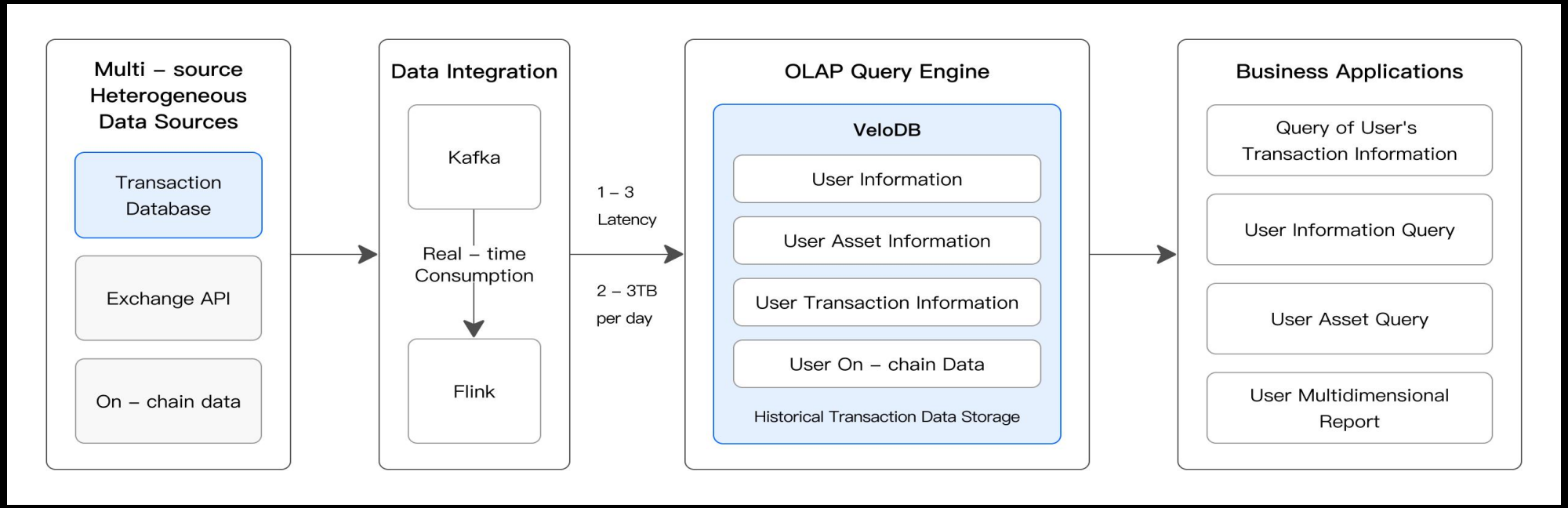
- •User Segmentation: Behavioral segmentation of participants for targeted re-engagement.
- •Business Promotion: Data-driven marketing campaigns based on competition insights

Current Status and Benefits

- Query Type: High-concurrency point lookup
- Core Business Scenario: Real-time leaderboard and user performance analysis during trading competitions
- Query Throughput: 1000+ QPS
- Latency (P95): ≤ 30ms
- Data Volume: Several terabytes (TB)

Use Cases for Web3

Scenario 5: User Trading Behavior Analysis



Business Scenarios

- User transaction lookup
- User profile inquiry
- User asset query
- Multi-dimensional user reporting

Application status and benefits

- Query Mode:Ad-Hoc detailed data analysis
- Core Use Cases:
  - Long-term historical transaction data storage
  - On-demand querying of user and transaction records
- Query Throughput: 200+ QPS
- P95 Latency: 1-3 seconds
- Data Volume: 400+ TB

Welcome to Doris Community

Subscribe

Subscribe to our mailing list and join our discussion

dev@doris.apache.org

Get technical support

- **Slack:** apachedoriscommunity.slack.com
- **Wechat Group**: Scan the QR code on the right.

Doris

Doris AI

Address Labels， Fund Flows & Graphs: Building a Crypto Risk Control System with NebulaGraph

Speaker: YunFei Xie
Security R&D Engineer, BlockSec
Date: August 22, 2025

Overview

1. The New Frontier: Crypto's Rise and Regulatory Challenges
2. The Core Idea: From Transaction Data to a Risk Network
3. The Implementation: Building the Risk Control System with Nebula Graph
4. A Case Study in Compliance: Anatomy of a High-Risk Transaction
5. The Value & Challenges

The New Frontier: Crypto's Rise and Regulatory Challenges

The Opening Act: A Market Too Big to Ignore

From Experiment to Global Financial Force:

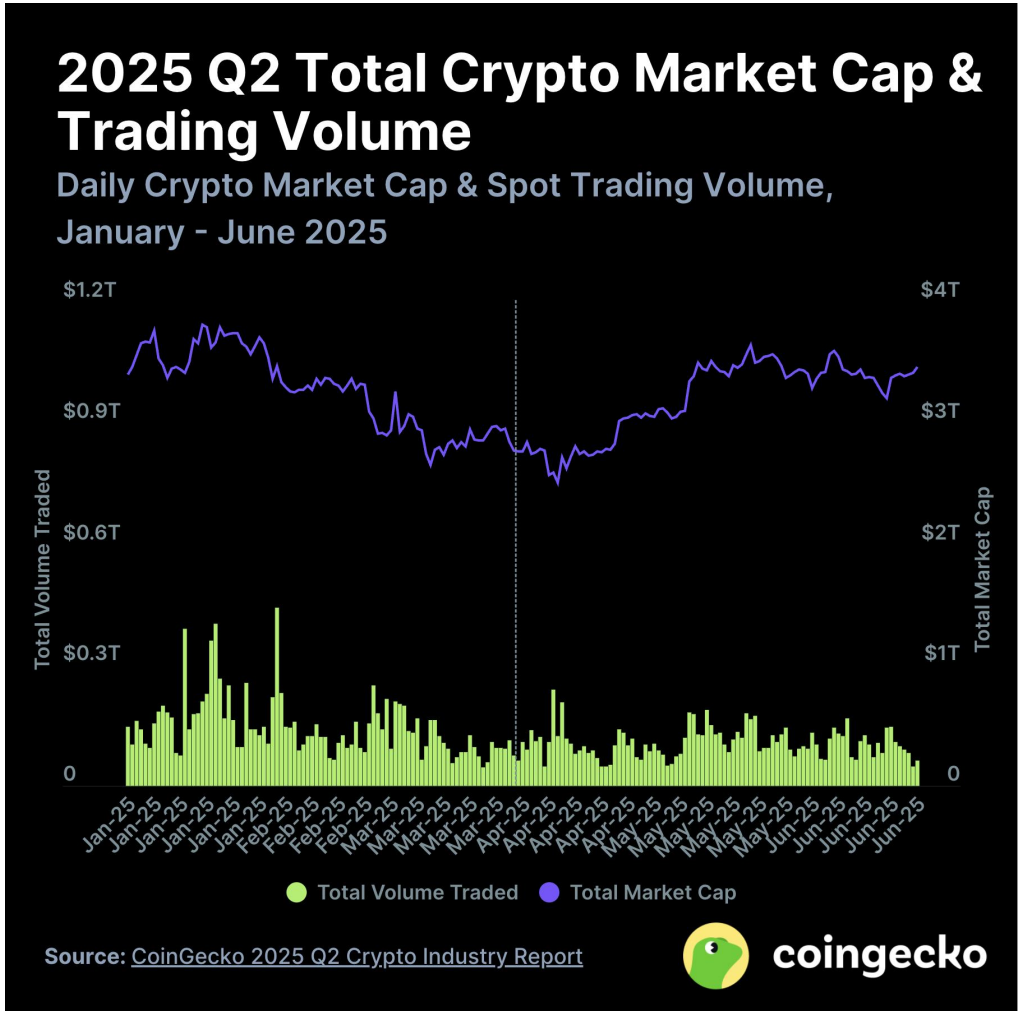
- **Rapid Evolution:** Crypto transformed from niche tech to a global financial power.
- **Market Scale (Mid-2025):** Market cap hit **$3T+**, nearly **3%** of global equity value.

Key Growth Metrics (H1 2025)

- **Global Users: 600M+** active users in the crypto ecosystem.
- **Trading Volume:** Daily volume now exceeds **$200B**.
- **Institutional Growth:** A massive 86% of institutions are now invested or plan to invest in 2025.

Implications for Global Finance

- **Unprecedented Speed:** Growth has outpaced the early internet.
- **New "Shadow Banking":** A parallel financial system has emerged.
- **Regulatory Challenge:** The new system tests traditional financial oversight.



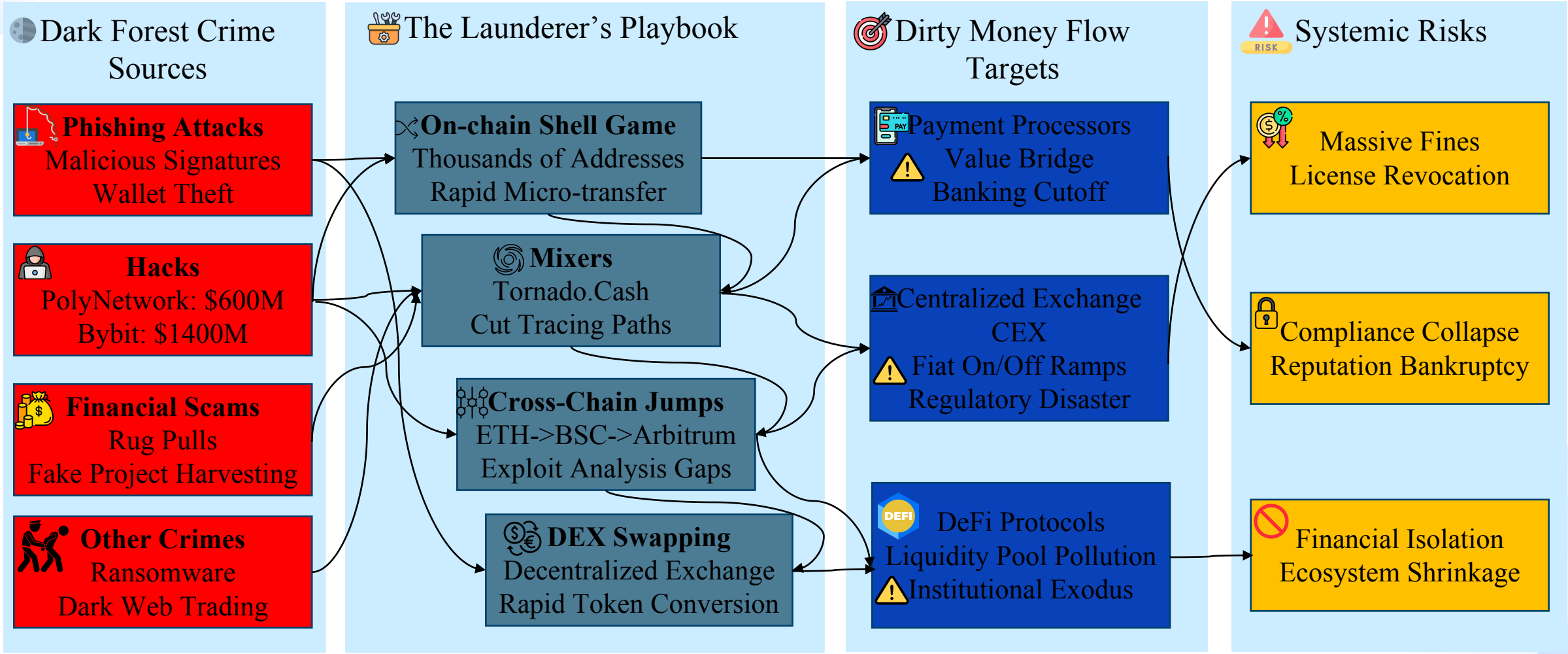
The New Frontier: Crypto's Rise and Regulatory Challenges

The Breakthrough: Stablecoins , Crypto's First Mass Adoption

1. From a Tainted Past	2. To Real World Utility	3. To a Global Financial Layer
Early impressions were dominated by illicit use cases, creating a significant perception hurdle.	Stablecoins gained traction by solving tangible problems traditional finance could not.	This utility has driven stablecoins to a scale rivaling.
Historic Stigmas:	**Killer Use Cases:**	**Key Metrics (2025 Data):**
Darknet Markets: Preferred currency for illegal trade (e.g., Silk Road). **Money Laundering:** Tools for illicit finance (e.g., Monero). **ICO Scams (2017-18):** >85% of projects failed or were fraudulent. **Ransomware:** The default payment demanded by cybercriminals.	**Hyperinflation Hedge (Argentina):** USDT adoption surged as a store of value against >95% inflation. **Efficient Remittance (Vietnam, Philippines):** Cross-border costs slashed from **6-8% to <1%**. **Financial Inclusion (Africa):** In key African markets like Nigeria, grassroots crypto adoption now rivals traditional finance	**Volume Rivals Payment Giants:** Stablecoins surpassed Visa and Mastercard combined transaction volumes in 2024. **Market Cap Hits All-Time High:** - Total: >$260 Billion - Dominance: USDT (62%)、USDC(25%)
Key Quote: *Crypto is nothing without illicit economy.*	**The Result:** *Building trust and utility beyond the crypto-native world.*	**The Conclusion:** *A new layer of the global payment infrastructure.*

The New Frontier: Crypto's Rise and Regulatory Challenges

The Shadow: A "Dark Forest" Before Regulation



The New Frontier: Crypto's Rise and Regulatory Challenges

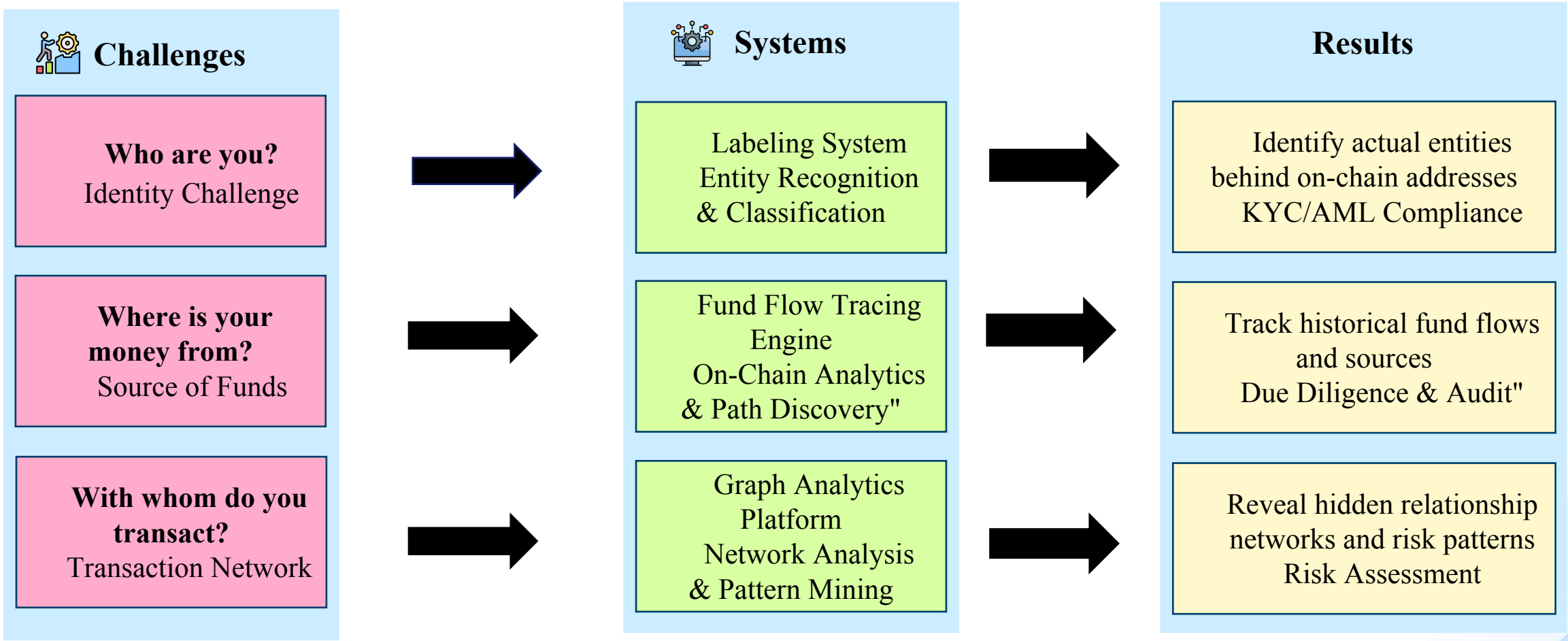
The Dawn of Compliance: Shaping the Global Landscape

FATF "Travel Rule"	MiCA	OFAC Sanctions	Hong Kong VASP License
Breaking Anonymity: Requires VASPs to share originator and beneficiary information during transfers, enabling regulators to trace the complete path of funds across platforms.	**Comprehensive Legal Framework:** Internalizes FATF recommendations into a binding EU law. Mandates ongoing transaction monitoring and dynamic customer risk rating systems.	**"Surgical Strike" Enforcement:** A "zero-tolerance" approach that adds non-compliant entities and addresses (like Tornado Cash) to the SDN blacklist.	**Compliance as Market Access:** The risk management system itself is a core component subject to regulatory review and approval for licensing.
KYC Binding: Link addresses to verified real-world identities. **Real-Time Vetting:** Assess risk before transactions commit.	**Pattern Detection:** Identify complex laundering patterns beyond simple rules. **Behavioral Scoring:** Use graph analytics for dynamic user risk scoring.	**Labels**: A powerful labeling system for real-time flagging of sanctioned addresses.	**Auditability:** Generate automated, regulator-ready compliance reports. **Investigations:** Provide visual tools to trace and analyze suspicious fund flows.

The New Frontier: Crypto's Rise and Regulatory Challenges

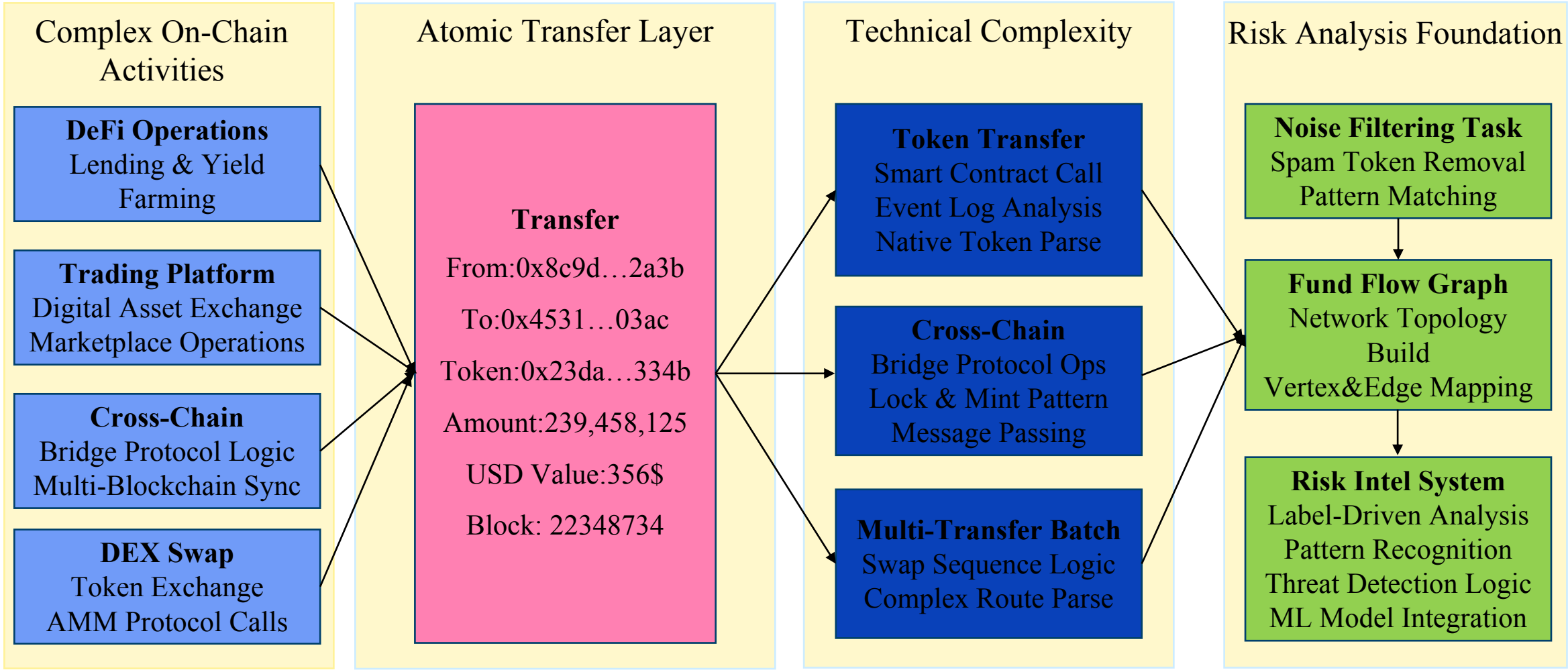
The Dawn of Compliance: Shaping the Global Landscape

Core Technical Challenges for Web3 Compliance The Three Pillars



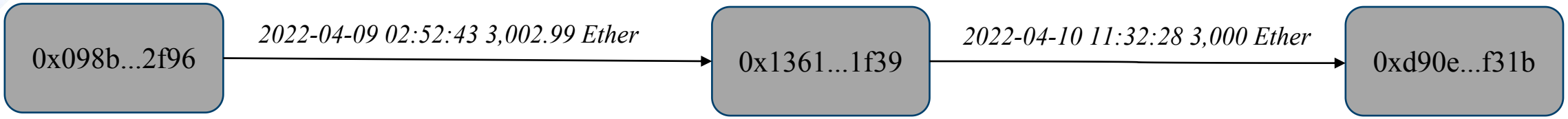
The Core Idea: From Transaction to a Risk Network

The Atomic Unit: The Transfer as the core of fund flow



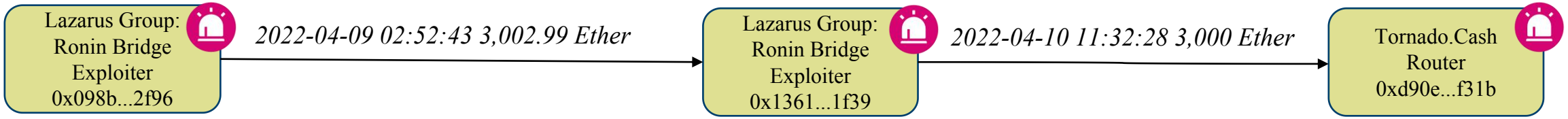
The Core Idea: From Transaction to a Risk Network

The Risk Network: The Graph is the Skeleton, Labels are the Soul



The Anonymous Skeleton: Limited Insight

- Raw Blockchain data presents us with anonymous address connections
- This creates a “skeleton” – structural connections without meaningful context



Labels Breathe Lift into Data

- Lazarus Group: Ronin Bridge Exploiter -> Tornado.Cash Router
- Suddenly, we see a 3000Ether hack being laundered through a sanctioned mixer

The Core Idea: From Transaction to a Risk Network

Illuminating the Threats: Precision Profiling of On-Chain Risks

Exploit	Phish	Pig-Butchering	Ransomware	Terrorist Financing
DeFi exploitation for gain	Social engineering & theft through deception	Fraudulent Investment via Emotional Grooming	Monetized cyber extortion	Covert funding via anonymity

DarkWeb Business	Sanctioned	Laundering	Mixer	Blocked
Underground crypto markets	Under official sanctions	Cleaning illicit funds	Obfuscating transactions	Blacklisted by notable contracts

The Implementation: Building the Risk Control System with Nebula Graph

System Architecture: The Four Pillars of the Risk Control System

- **Data Layer**

Real-time, filtered on-chain fund flow data extraction and processing

- **Intelligence Layer**

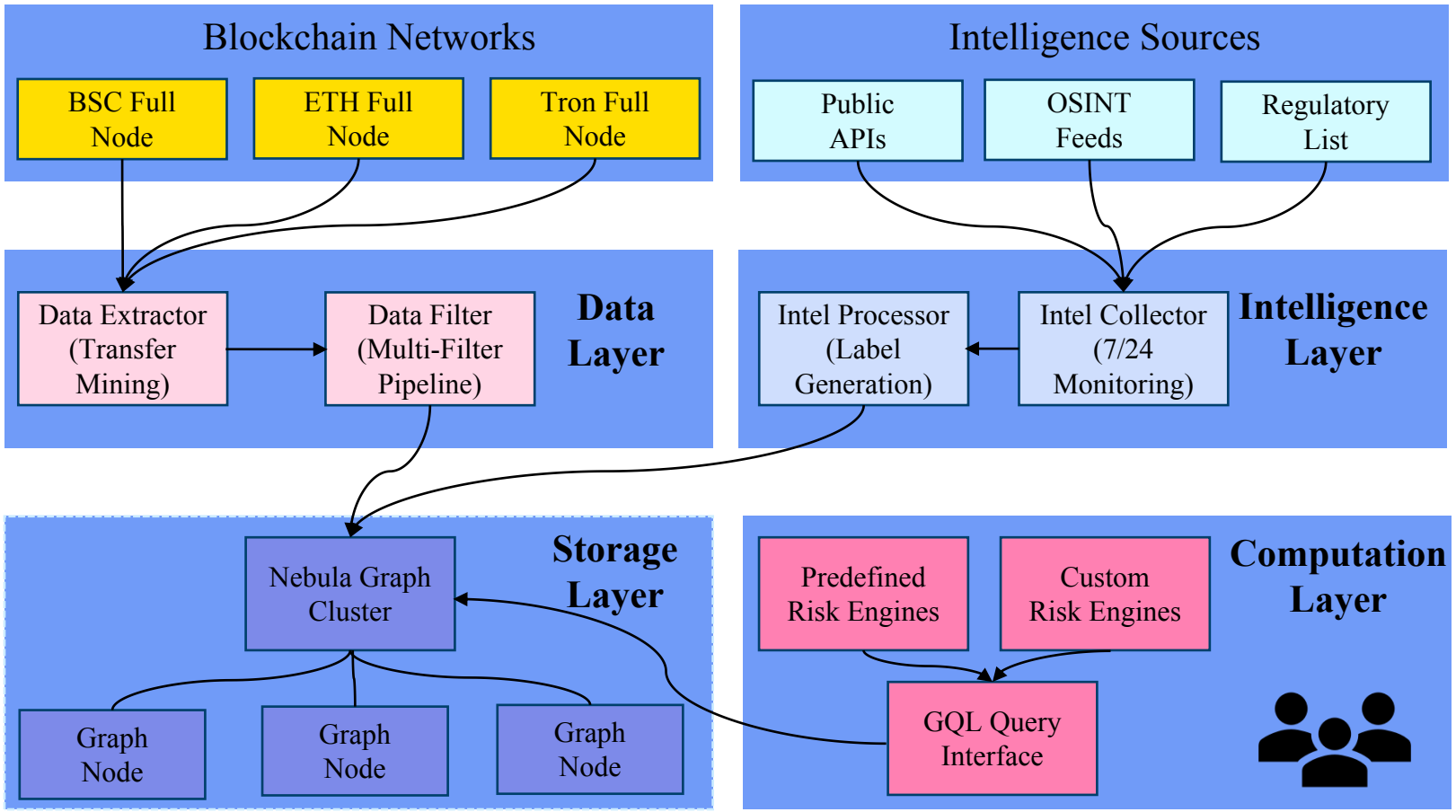
Precise and accurate address labeling system from multiple intelligence sources

- **Storage Layer**

High-performance graph database powered by Nebula Graph

- **Computation Layer**

Advanced risk engines for pattern detection and threat assessment



The Implementation: Building the Risk Control System with Nebula Graph

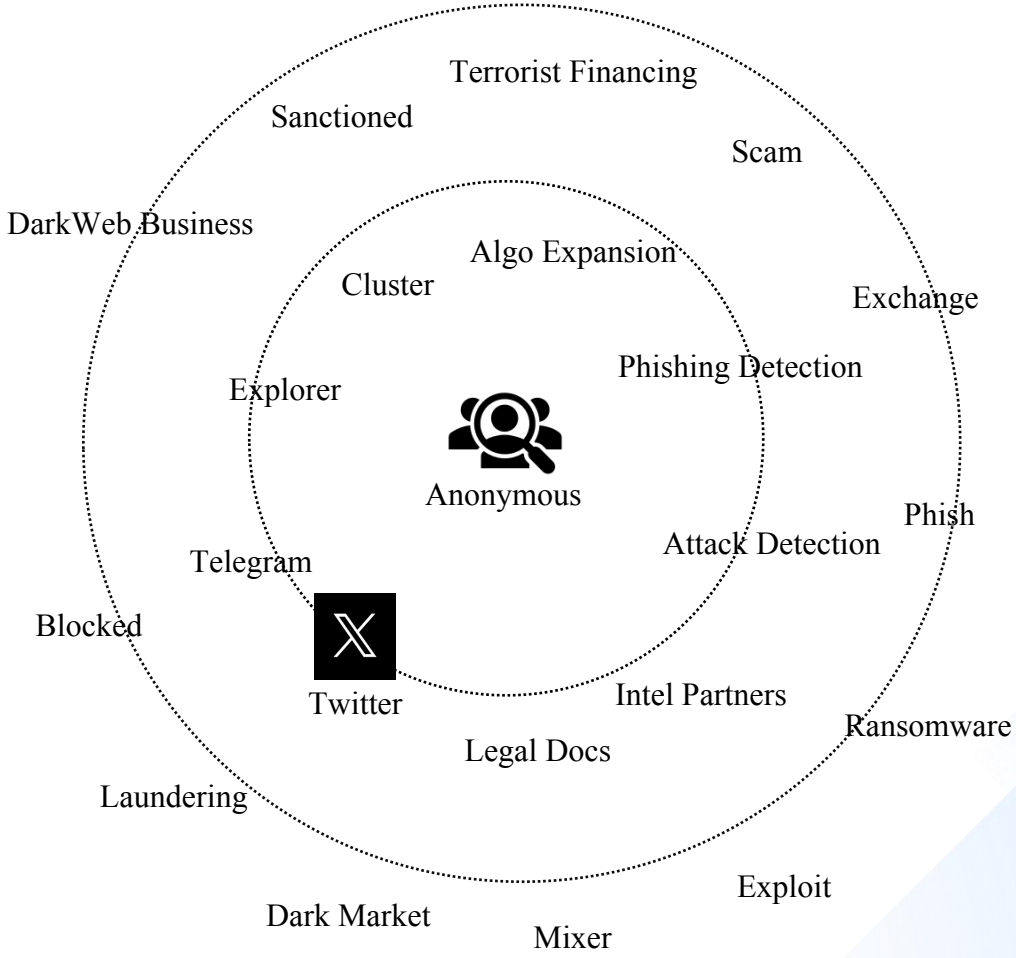
Qualitative Analysis: Using Labels to Give Data Meaning

Where Do Labels Come From?

- **On-Chain Heuristics:** Behavioral patterns from blockchain activity.
- **Off-Chain Intelligence:** Proactive collection from public sources (Twitter, Telegram), official legal documents, third-party intelligence partners.
- **Human Analysis:** Expert-driven investigation to verify automated findings and uncover novel intelligence that systems may miss.

How to Ensure Label Accuracy?

- **Automated Cross-Validation:** Our system automatically aggregates and cross-references data from multiple intelligence sources
- **Human-in-the-Loop Feedback:** We leverage a continuous feedback cycle from our suite of products (MetaSleuth, Phalcon, API services).



The Implementation: Building the Risk Control System with Nebula Graph

Risk Quantification: Building a Computable Risk Model

Answering the Question: How Much Funding is from Scammers vs. Exploiters?

To quantify the risk and trace the sources of funds, we apply a heuristic algorithm with the following key steps:

1. Fiat Exposure Calculation

- **Objective:** Standardize the USD value of token transfers.
- **Approach:** Use **price oracles** to calculate the USD equivalent for every token transfer. This allows us to compare risks and contributions across different assets consistently.

2. Time Series Filtering

- **Objective:** Ensure temporal consistency in tracing.
- **Approach:** Only edges in compliance with a time sequence rule are preserved. An address cannot spend funds that it has not yet received.

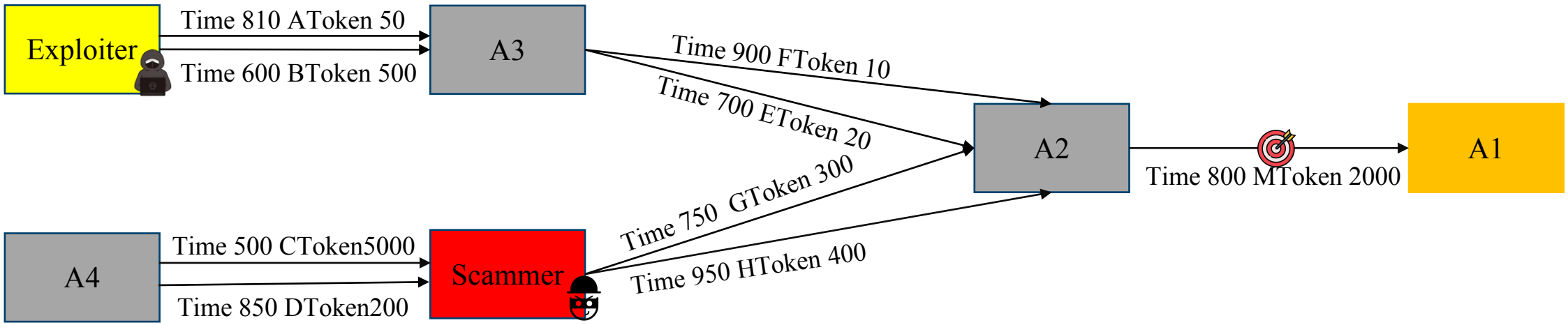
3. Haircut Strategy

- **Objective:** Focus on the most relevant paths.
- **Approach:** Apply proportional weight allocation and introduce a **threshold** to prune low-contribution or irrelevant paths. This prevents contamination of the result by negligible or noisy funding sources.

A Case Study in Compliance: Anatomy of a High-Risk Transaction

Understanding the Flow Diagram:

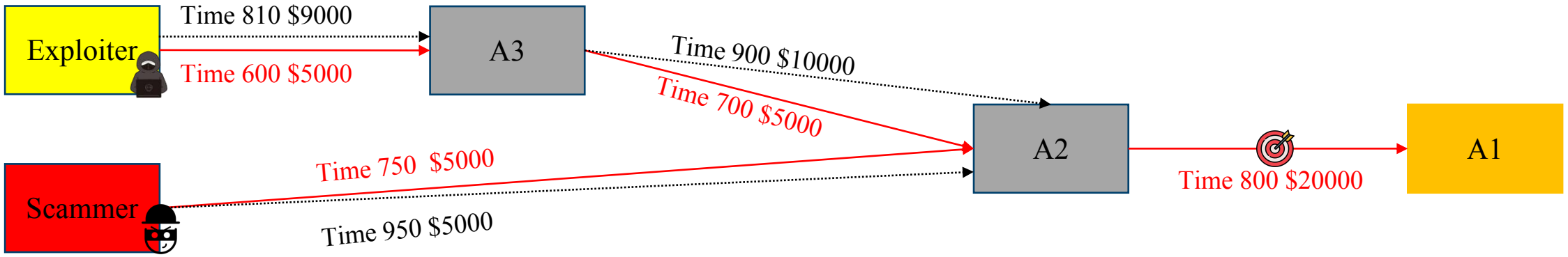
- **Rectangles** represent addresses in the network
- **Arrows** represent token transfers between addresses
- **Time values**: Higher numbers indicate later timestamps
- **Token amounts**: Numbers after token symbols represent transfer amounts



Core Question:

How much of the funds in the A2→A1 transfer originated from the Exploiter versus the Scammer?

A Case Study in Compliance: Anatomy of a High-Risk Transaction



Step 1: Identify Valid Time-Filtered Paths (Red Paths Only)
Time Reference Point: A2→A1 transfer occurs at **Time 800**
Valid Path Analysis (Red Paths):

Exploiter → A3: $5,000 (Time 600) → A3 → A2: $5,000 (Time 700)

Scammer → A2: $5,000 (Time 750)

Exploiter → A3: $9,000 (Time 810) - **INVALID** (occurs after Time 800)

A3 → A2: $10,000 (Time 900) - **INVALID** (occurs after Time 800)

Step 2: Proportional Allocation of A2→A1 Transfer
A2→A1 Transfer Amount: $20,000 (Time 800)
Since A2 only has $10,000 in valid inflows but transfers $20,000, we calculate proportional contamination:
Fund Source Proportions:
From Exploiter: $5,000 ÷ $10,000 = 50%
From Scammer: $5,000 ÷ $10,000 = 50%
Allocation in $20,000 Transfer:
From Exploiter: $20,000 × 50% = **$10,000**
From Scammer: $20,000 × 50% = **$10,000**

The Value & Challenges

Proactive Prevention	In-Process Control	Post-Mortem Forensics
Building a Proactive Defense System	**Raising the Security Baseline of Partners**	**Providing Precise Investigation & Evidence**
Core Capability: Real-time **blocking** of risky transactions.	**Core Capability:** Comprehensive **due diligence** on counterparties and projects.	**Core Capability:** Deep **penetration and tracing** of on-chain funds.
Application Scenario: *When processing user deposits on an exchange or interactions on a DeFi protocol, an API call can fetch the risk score of associated addresses. If the score exceeds a threshold (e.g., funds from a mixer or sanctioned entity), the system can **automatically halt or reject the transaction**, stopping the risk at the gate.*	**Application Scenario:** *When evaluating a new DeFi project for integration or conducting a large transaction with an institution, graph analysis can review the historical behavior and funding network of their contracts and wallets. This effectively identifies if they have close ties to **"high-risk entities,"** providing critical data for partnership decisions.*	**Application Scenario:** *After a security incident (like a hack), provide clear, visual fund flow reports for the project team and law enforcement. Our system can quickly trace the complete path of stolen funds through multiple hops, mixers, and into exchanges, securing the **crucial time window for asset freezing and recovery**.*

The Value & Challenges

The Data Challenge	The Complexity Challenge	The Adversarial Challenge
Label Quality & Breadth Define the System's Ceiling	**Cross-Chain and Privacy Tech**	**The Continuous "Arms Race" with Illicit Actors**
Accuracy is the Lifeline: A single incorrect label can lead to catastrophic misjudgments. The generation and verification of labels demand an **extremely rigorous process**.	**"Breakpoints" in Cross-Chain Tracking:** When funds move between blockchains via bridges, it creates analytical "breakpoints." **Effectively connecting these disparate graphs** is a significant technical challenge.	**"The Devil is Always a Step Ahead":** As we refine our tracking algorithms, illicit actors upgrade their anti-tracking methods, such as using DEXs for **rapid, multi-path asset swaps** or new, untagged mixing protocols to evade analysis.
Coverage is the Ceiling: *Facing a massive and growing number of new addresses daily, the ability to quickly and effectively label them is the fundamental limit of our system's reach.* ***Vast "unknown territories" will always exist.***	**Adapting to Emerging Ecosystems:** *The constant emergence of Layer 2 solutions and new public chains, each with unique data structures, requires our system to have* ***extreme scalability and rapid adaptation capabilities.***	**A Never-Ending Race:** *The risk control system must* ***constantly iterate and evolve*** *to maintain an edge in the ongoing confrontation with malicious actors.*

Thank You

Q&A

https://nebula-graph.com.cn

GitHub: vesoft-inc/nebula

Twitter: @NebulaGraph

Facebook: @NebulaGraph

https://discuss.nebula-graph.com.cn

微信公众号

开源项目

NebulaGraph | VELODB

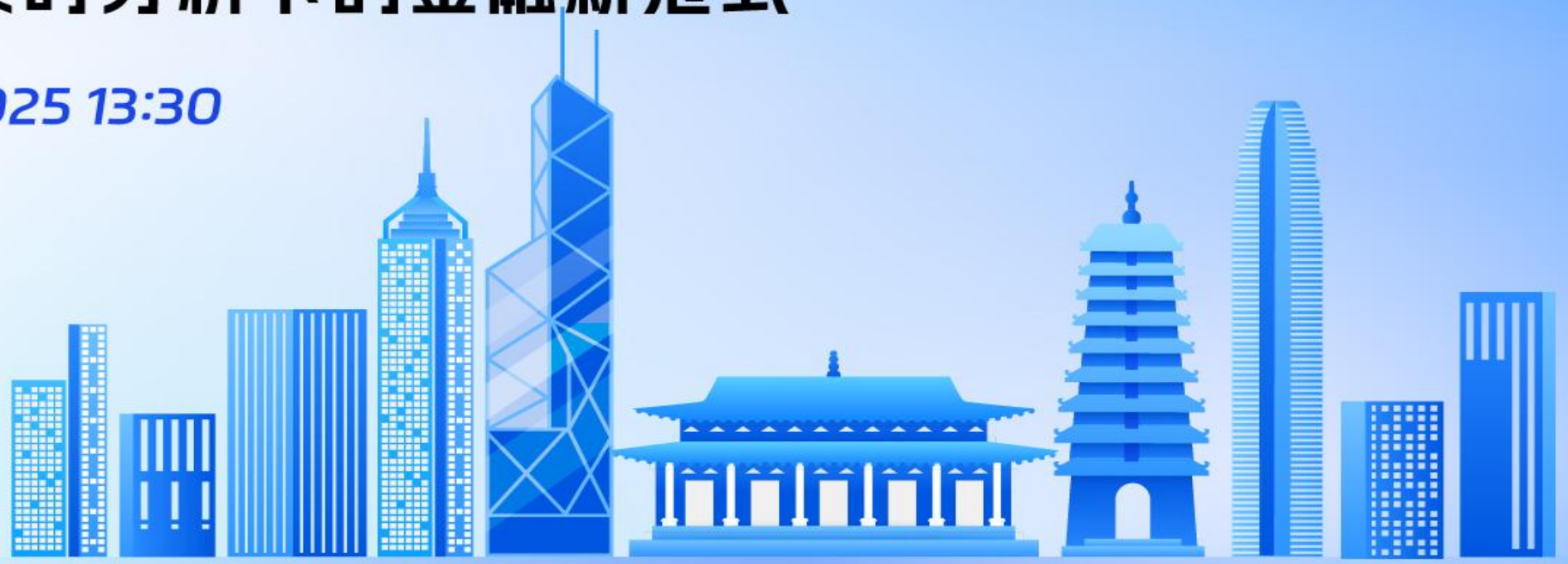
Deep Relationships & Real-Time Analytics:

Architecting Modern Financial Data Stacks

深度关联与实时分析下的金融新范式

August 22, 2025 13:30

Hong Kong



普适智能
PUSHI AI

Knowledge Graph-Based Web3 Risk Control Strategies

PUSHIAI Founder: Chen Xiaoxiang

Leader in Graph Computing Large Models × Privacy-Preserving Computation × XinChuang Ecosystem

contents

PART ONE About Us

PART TWO Knowledge Graph-Based Web3 Risk Control Strategies

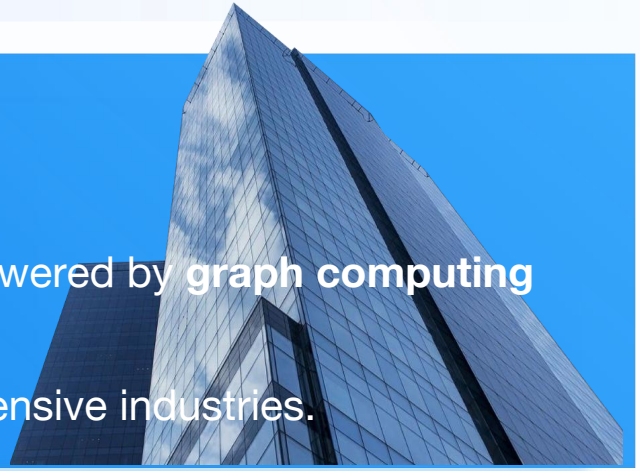
PART THREE Product Overview

Delivering Value Through AI-Driven Industrial Decision-Making

PUSHI AI

Specializing in Graph Computing Large Models · Building a Federated Application Ecosystem

Company Snapshot



- Founded in 2018 | HQ: Shenzhen
- **Mission**: Deliver measurable business outcomes through intelligent decision engines powered by **graph computing large models**, **integrating privacy-preserving computation** and **big data**.
- **Key Value**: Transition from "decision support" to **automated value creation** in data-intensive industries.

Key Milestones

- **2015-2017**: Prototyped graph computing engines in Fortune 500 R&D labs.
- **2018-2019**: Launched 1st-gen platform, partnered with top-tier banks.
- **2020-2021**: Validated cross-industry scalability with 3rd-gen dynamic graph engine.
- **2022-Present**: Served 30+ institutions nationwide; expanding globally via Huawei and XinChuang ecosystem alliances.

Delivering Value Through AI-Driven Industrial Decision-Making

Specializing in Graph Computing Large Models · Building a Federated Application Ecosystem

Key Achievements

•**Industry Applications**: Deployed in financial regulation (anti-money laundering, fraud detection), energy management (smart power station decision systems), and public safety (intelligent case analysis).

•**Core Team**: Composed of AI experts and industry veterans from Fortune 500 companies, deeply committed to translating AI into tangible results.

Future Vision

Integrating graph computing large models into a **"smart decision OS"** and AI agent ecosystems to drive autonomous, industry-wide intelligent upgrades.

Why It Matters

•**Core Team**: Fortune 500 AI experts + industry veterans.

•**Tech Edge**: Seamlessly integrates graph computing, privacy tech, and federated learning.

•**Outcome-Driven**: Directly links AI to ROI (e.g., KPI improvement, profit growth).

Knowledge Graph Technology

PUSHI AI

Current Challenges： **Overreliance on Traditional Tools** - Heavy dependence on big data and machine learning, yet constrained by labeled data requirements and limited ability to mine multi-hop relationship networks.

		Key Requirements	Limitations of Traditional Tech
Risk Management	**Escalating Organized Fraud**: Sophisticated syndicated fraud, such as credit card and loan scams, evades traditional rule-based models designed for isolated entities. **Hidden Risk Propagation**: Risks spread through complex networks (e.g., corporate groups, supply chains).	Advanced relationship network analytics (multi-layer, dynamic links) Real-time risk propagation tracking Intelligent detection of fraud rings/syndicates	**Performance Bottlenecks**: Queries with ≥3 relationship layers suffer latency spikes. **High Development Costs**: Custom coding for diverse data sources and rules. **Implementation Gaps**: No efficient tools for fraud ring detection or dynamic network mining.
Targeted Marketing	**Low Conversion of Customer Insights**: "Referral-based" strategies lack effective outreach channels.		
AML	**Obfuscated Transaction Networks**: Requires dynamic analysis of multi-layer fund flows to uncover laundering patterns.		

Technical Features of Knowledge Graph

PUSHI AI

Knowledge graph technology employs graph databases (vertices/edges) to build dynamic semantic networks, enabling real-time relationship analysis and flexible data modeling beyond relational table constraints.

The node-edge architecture inherently serves as a semantic framework, enabling **unified standard of business data life circle** through structural integration and contextual chaining.

Knowledge graphs leverage graph-based algorithms—including topological analysis, shortest path optimization, Louvain community detection, and connected component identification—to transform data into relational networks, significantly **enhancing analytical capabilities for group detection.**

Graph features of individual vertices address the limitations of traditional machine learning models, which focus solely on isolated statistical attributes, by incorporating relational contexts between entities into model training. This approach significantly enhances generalization capabilities.

Graph insights are visualized through node-edge networks, enhancing **clarity** and **explainability.**

contents

PART ONE About Us

PART TWO Knowledge Graph-Based Web3 Risk Control Strategies

PART THREE Product Overview

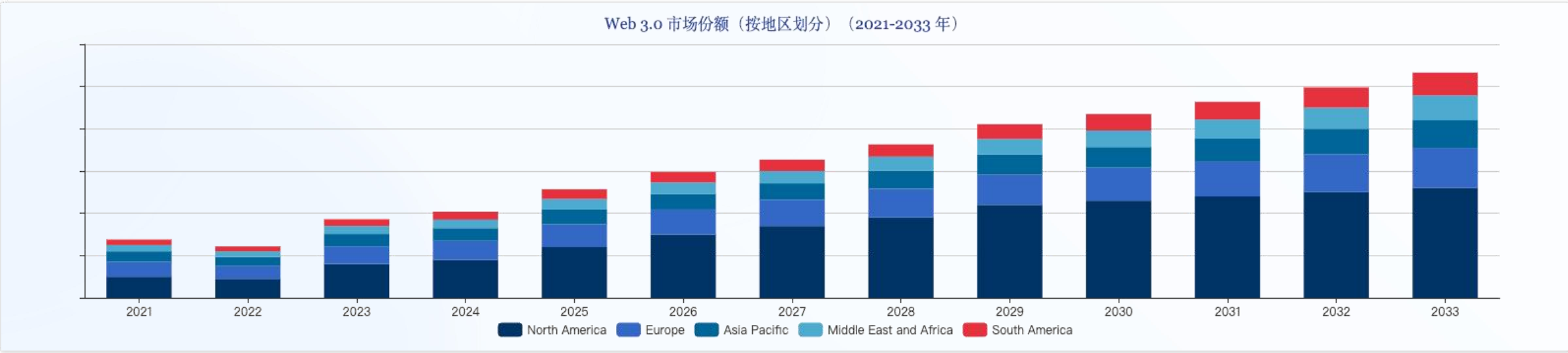
Digital Currency Exchanges: Evolution & Challenges

Global Web3 Landscape: The Decentralized Evolution

Web3 signifies a transformative shift toward democratized digital ecosystems, delivering enhanced security, privacy, and user sovereignty beyond Web2 limitations.

Market Insight

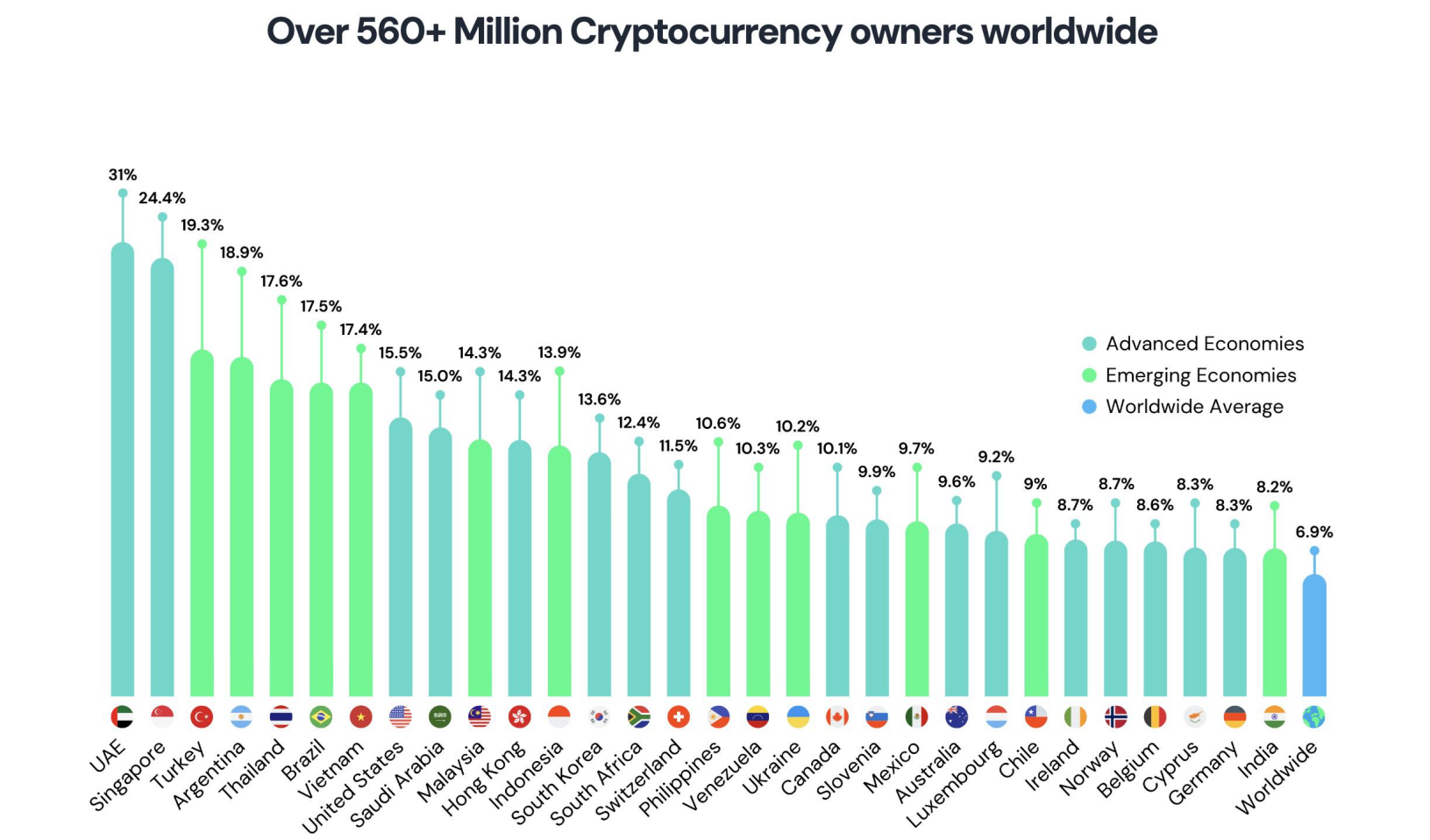
The global Web3 market is projected to surge from **$0.4B in 2023 to $5.5B by 2030** at a **44.9% CAGR** (MarketsandMarkets).



Digital Currency Exchanges: Evolution & Challenges

Global Digital Currency Evolution

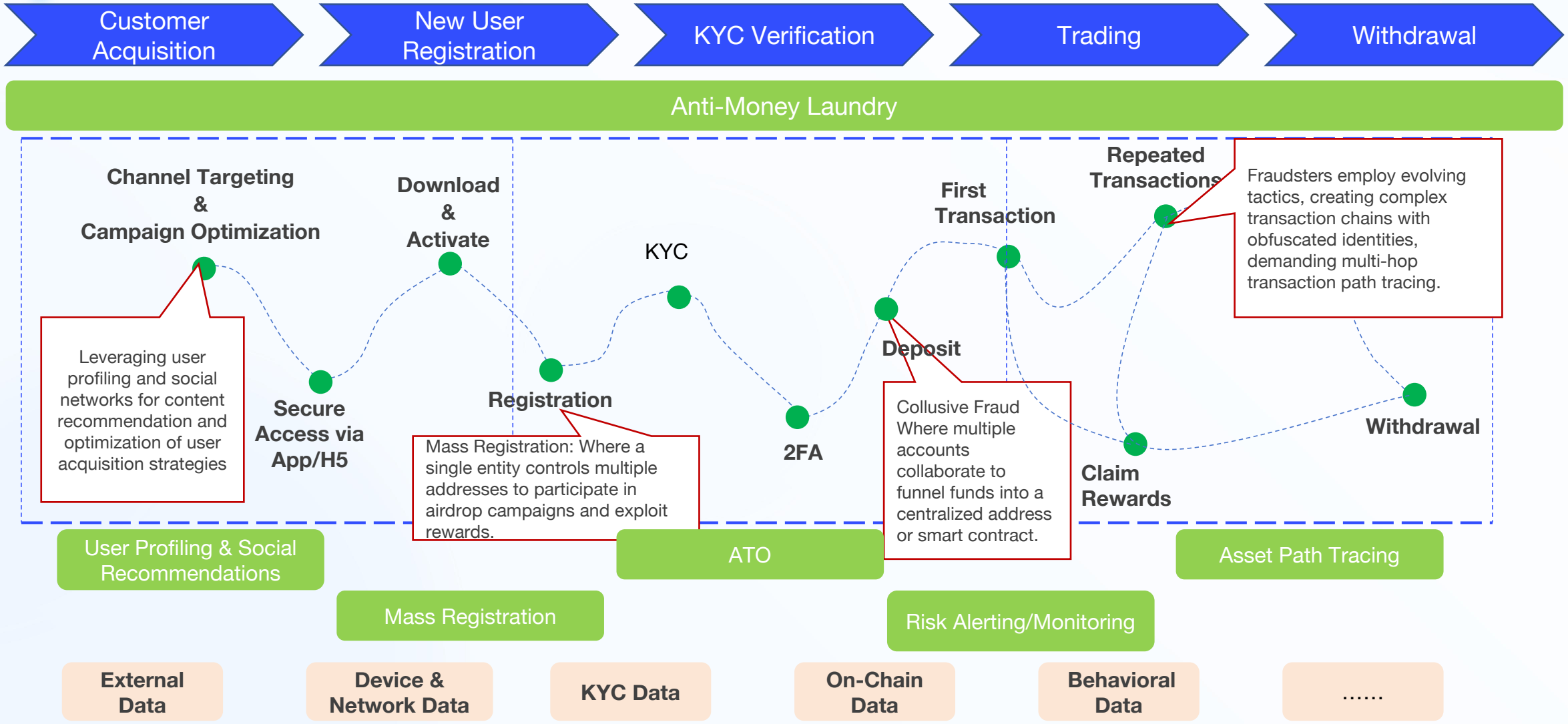
Global digital currency trends highlight leading native tokens—BTC, ETH, ADA, TRX, and ATOM—while revealing accelerated worldwide adoption, with current ownership reaching 6.8% globally (560+ million users), marking a 33% year-on-year surge from 2023's 420 million users; notably, 34% of holders are aged 24-35, and the UAE demonstrates the fastest growth at 25.3% according to Triple-A's 2024 Global Crypto Ownership Report.



Digital Currency Exchanges: Evolution & Challenges

PUSHI AI

Graph Database Applications in Digital Currency Exchange Business



Digital Currency Exchanges: Evolution & Challenges

PUSHI AI

The core challenges faced by digital currency exchanges under traditional technology systems

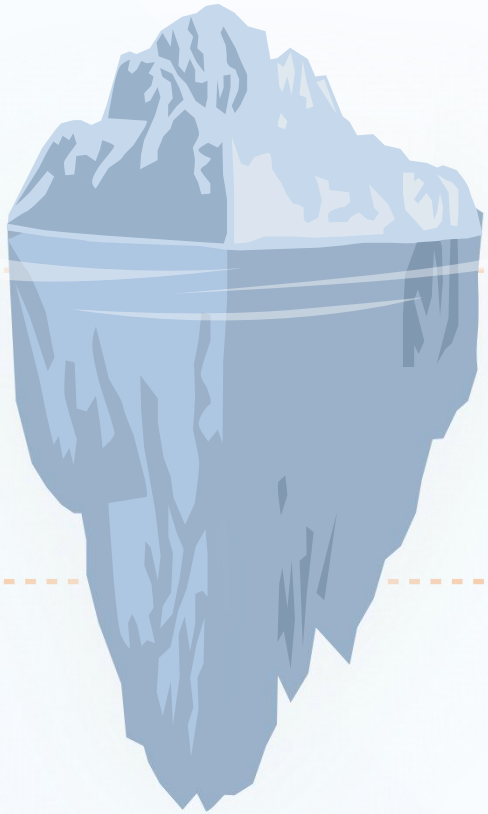
Digital currency exchanges demand relationship-intensive, path-dependent analysis, while traditional table-centric systems face bottlenecks in modeling, querying, computation, and visualization, hindering complex entity relationship mining and risk insights.

1 Technical Challenges in Integrating Billion-Scale On-Chain and Off-Chain Data"

Cross-chain transactions, contracts, and assets form massive dynamic graphs, while fragmented exchange data impedes efficient on-chain/off-chain address-to-user mapping, constraining analytical depth and scope.

2 Challenges in Tracking Criminal Networks' Identity Masking and Deep Fund Flows

Criminal networks conceal beneficiaries and fund flows through multi-address per user, address sharing, cross-chain hopping, and transaction layering. Traditional account-centric modeling fails to expose hidden links and money trails, escalating AML and compliance pressures.



3 High-concurrency scenarios intensify real-time risk control response challenges

Confronting emergent risks like HFT, arbitrage bots, and mass withdrawals, especially during mass registrations or rapid fund movements, traditional systems relying on rule engines and offline analytics struggle with millisecond-level detection and real-time blocking, leading to delays and false negatives in dynamic risk landscapes.

Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Architecture Landscape: Graph-Intelligence Solutions for Digital Currency Exchanges

Business Objectives

Resilient risk control and compliance, safeguarding sustainable growth

Business Workflow

New User Registration → KYC → Trading → Withdrawal

Application Scenarios

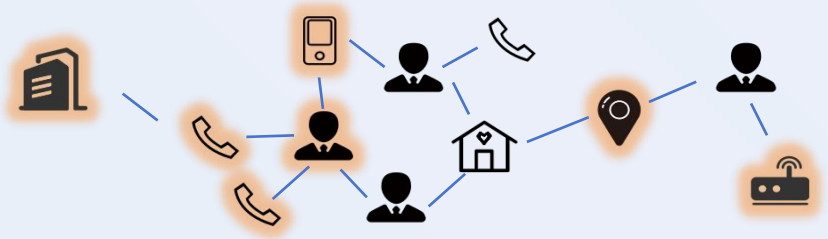
New User Registration	KYC	Trading	Withdrawal
Batch registration detection	Multi-account control analysis	Circular arbitrage pattern detection	Anomalous Large/High-Frequency withdrawal
Real-time account hijacking alerts	Detection of Hidden User Associations	Real-time fund flow tracking	Fund Tracing
Criminal account identification	Risk account penetration	Coordinated trading identification	Cross-Border Fund Flow Tracking

Graph Technology

- Real-time Decision Making
- Collusive Network Detection
- Data Exploration Capabilities
- Graph Features
- Graph Metrics
- Graph Algorithms
- Hidden Chain Detection
- Loop Pattern Detection
- Degree Centrality Analysis
- Community Detection Algorithms
- Similarity Algorithms

Graph Data Model

Build a Relationship Backbone



Data Layer

- External Data Sources
- Device & Network Fingerprints
- KYC Identity Vault
- On-chain Transaction Data
- Behavioral Telemetry
- .....

Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

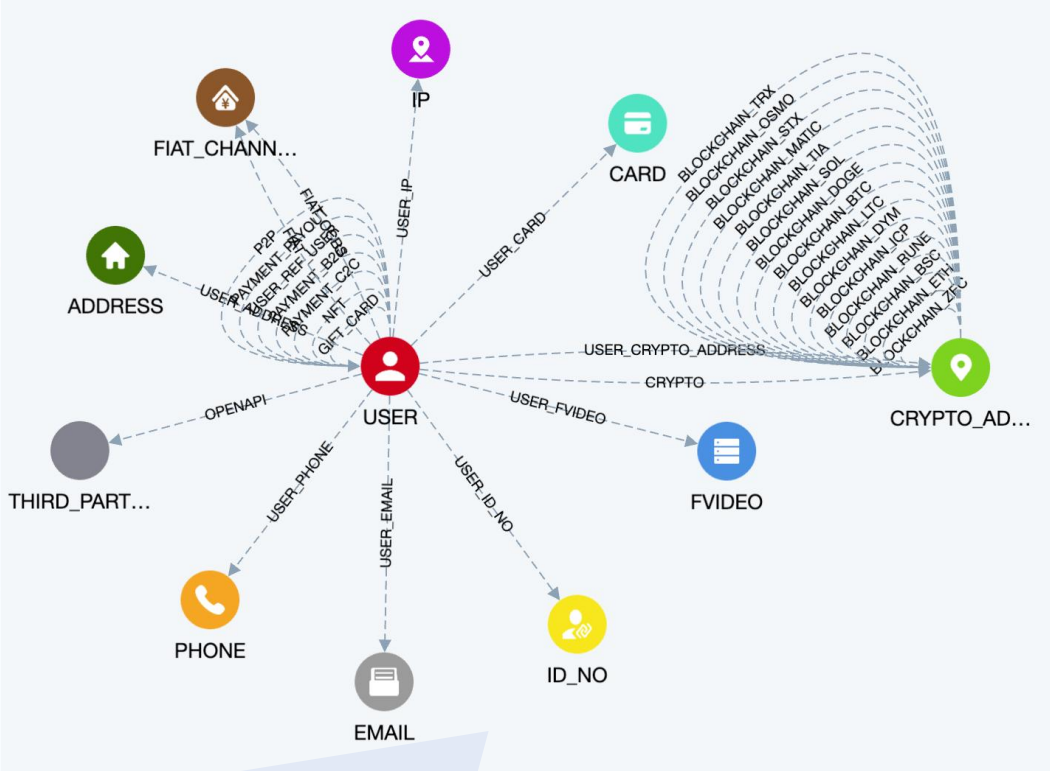
Constructing Graph-Based Models Using Digital Currency Exchange Data

Integrated multi-source data fusion enhances network connectivity, enriches entity attributes, and establishes a robust data foundation for relationship network detection.

- On-Chain Transaction Data
- Device & Session Fingerprints
- User Acquisition Records
- Web3 Social Graph Data
- ……

Batch Data Ingestion

Streaming Data Integration

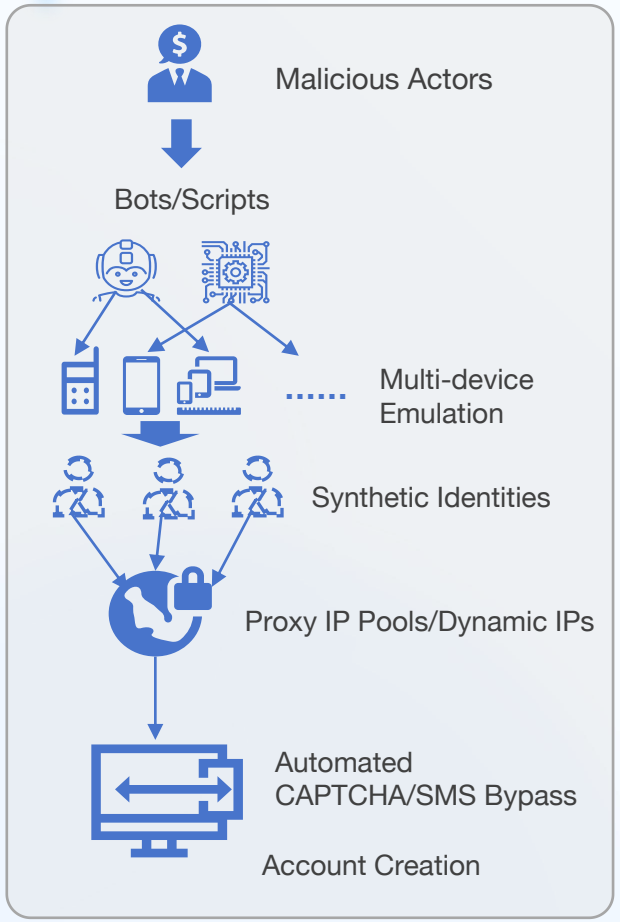


Integrates blockchain data (e.g., CRYPTO_ADDRESS, BLOCKCHAIN_XXX transactions) with in-house user data (e.g., ID_NO, EMAIL, PHONE, CARD, FIAT_CHANNEL) and capital flow data (e.g., P2P, PAYMENT_B2C), plus multi-channel user behavior data (e.g., IP, FVIDEO).

Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Traditional Mass Registration Risk Analysis



Generated Data Signatures

Data Signatures	Feature Footprint
Registration Info	Batch-generated emails/phones; Reused ID_NO/bank cards
IP Address	Rapid IP changes; Shared proxy IP ranges
Device Metrics	Identical FVIDEO; Near-identical browser fingerprints
Timing Patterns	Sub-second registration intervals; Fixed time-delta clusters
Behavior Path	Scripted paths: Registration → KYC → First deposit → Identical trades

Behavior-Based Risk Control Framework

Static Rules

- Single-IP registration volume
- Blacklisted address connections
- FVIDEO reuse threshold
- ……

Fingerprint ID

- Device/browser uniqueness validation

Blacklisting

- IP/FVIDEO/email domain denylists

Rate Limiting

- Registration request throttling

Critical Challenges in Legacy Systems

- **Siloed Dimension Analysis:** *Inability to dynamically correlate IP/FVIDEO/KYC/behavioral data*
- **Reactive Detection Lag:** *Rule-triggered post-event blocking fails real-time prevention*
- **Undetected Sybil Clusters :** *Traditional methods miss single operators behind puppet accounts*

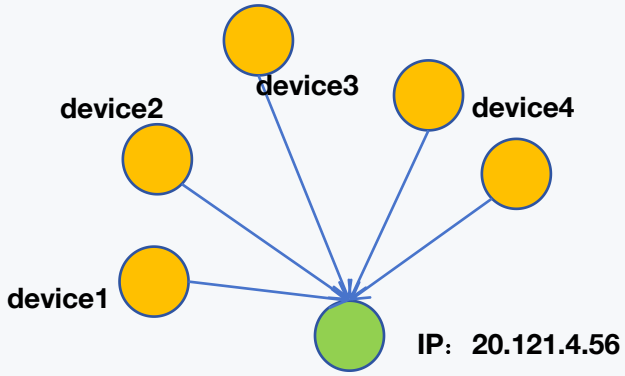
Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Mass Registration Defense: Real-time multi-dimensional correlation via graph computing

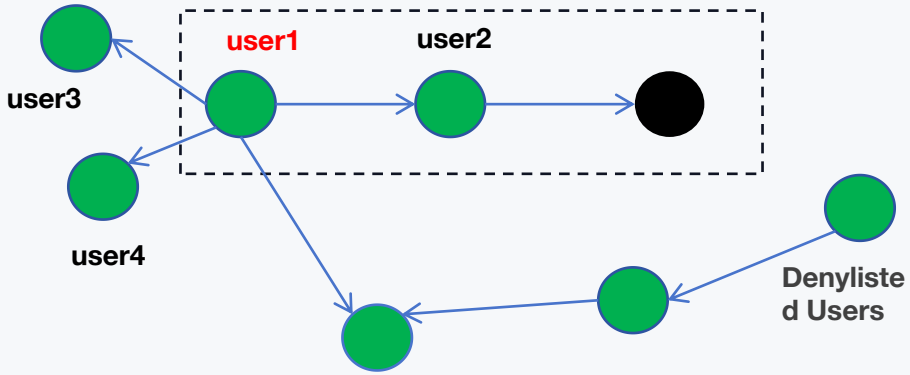
Graph-Rules Engine: Accelerating visual fraud logic development with real-time discrepancy checks and cross-dimensional correlation integrated into risk decision-making.

Sybil registration devices linked to shared IPs



Graph models uncover hidden relationships across IPs, FVIDEO, KYC, and behavior trails – exposing fraud rings and critical risks like IP sharing clusters and cross-account device reuse.

3-Hop Proximity to Blacklisted Addresses



Real-time structural matching against high-risk entity graphs during registration – dynamically correlating IP/FVIDEO/PHONE/EMAIL with denylisted nodes.

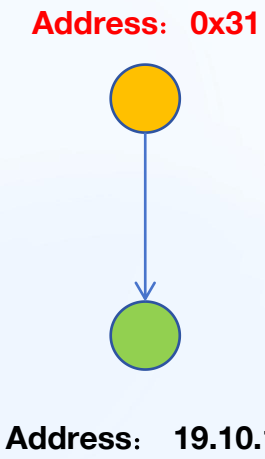
Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Mass Registration Defense: Real-time risk identification and blocking via graph topology analysis

Individual Address: Normal Status

Account ID	Registration Timestamp	On-chain Address	IP Address
1	T	0x31	19.10.12.4

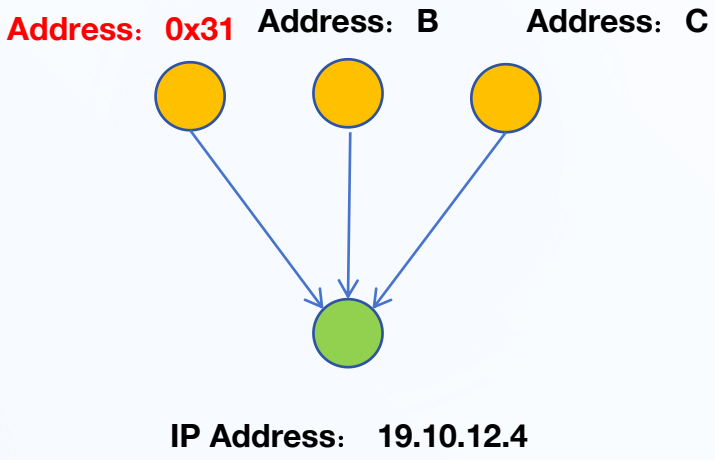


Quality Verified Accounts

Release Authorization
Rewards Distribution

Identify suspicious batches on ingestion , applying graph rules to globally contextualized, real-time writes.

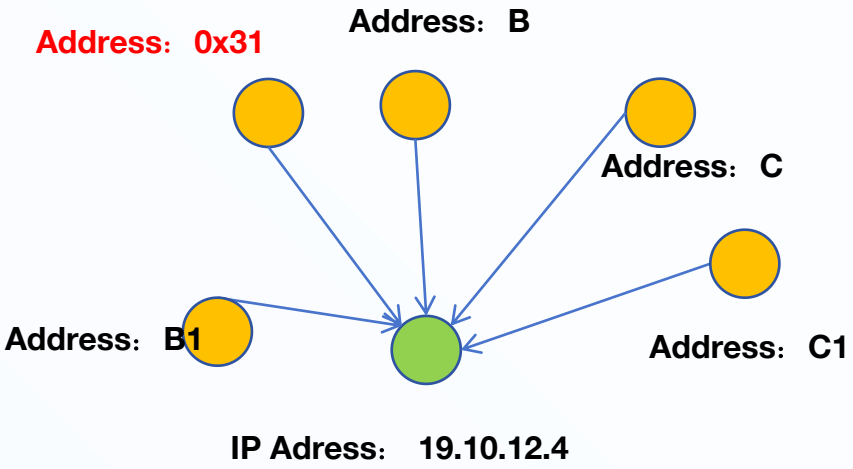
Account ID	Registration Timestamp	On-chain Address	IP Address
2	T+1	B	19.10.12.4
3	T+1	C	19.10.12.4



Flagged Fraudulent Accounts

Restrict Account Functions
Block Reward Payouts

Account ID	Registration Timestamp	On-chain Address	IP Address
4	T+3	C1	19.10.12.4
5	T+3	B1	19.10.12.4



Denylisted Fraud Accounts

Flagged High-Risk Addresses
Freeze Funds & Restrict Withdrawals

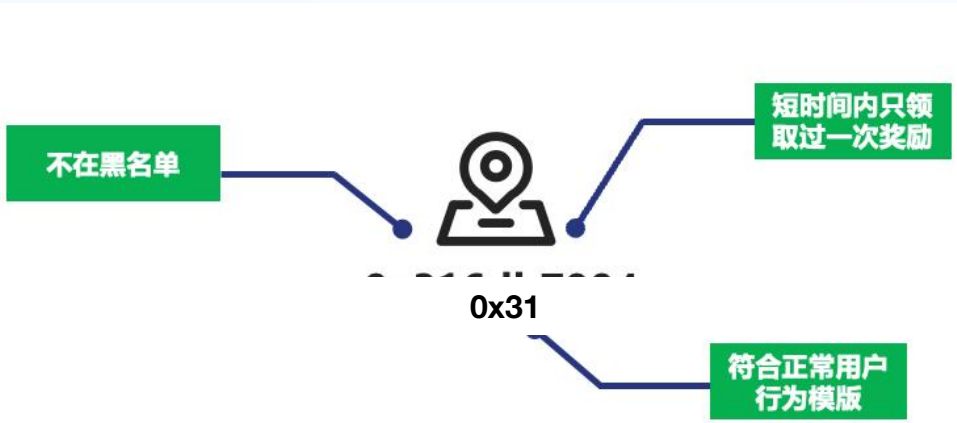
Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Mass Registration Defense: Real-Time Identification of Fraud Rings via Graph Topology Analysis

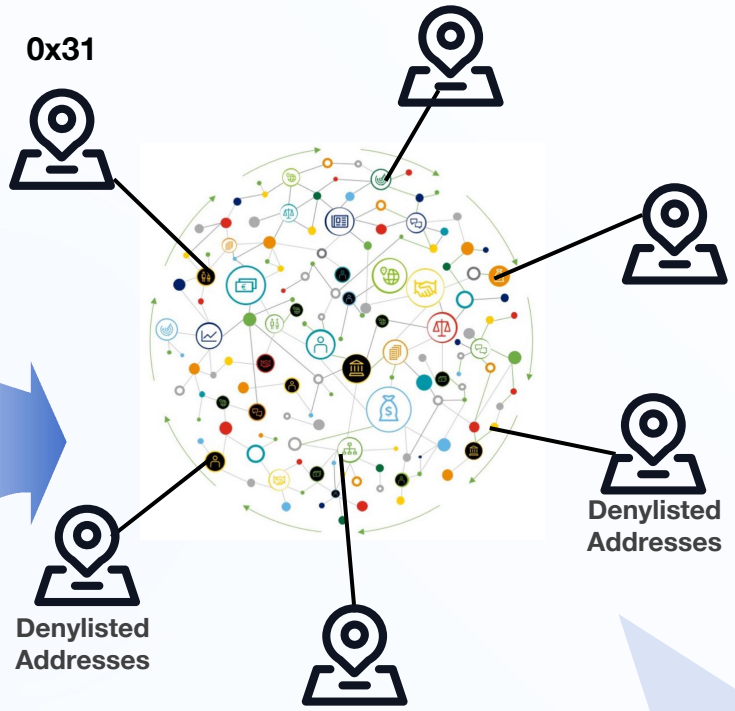
1.Complex Relationship Detection Failure
Fraudulent networks exhibit intricate fund flows & coordinated actions. Traditional rule-based methods miss hidden links.

2. Cross-Platform Risk Blind Spots
Massive transactional scale enables multi-exchange fraud. Single-point analysis cannot trace cross-platform risks.



Single-Point Risk-Free Assessment

Quality Verified Account Addresses



High-Risk Cluster Identification

Risk-Flagged Addresses

Group Behavior Rules

Batch Operations in Time Window
Multiple addresses execute bulk operations simultaneously within a cluster.

Fund Convergence to Single Node
Cluster funds ultimately flow to one central address.

Low External Connectivity Ratio
External connections of cluster < 20%.

High-Betweenness Centrality Nodes
Multiple nodes with elevated betweenness centrality exist.

High Internal Connection Density
Average intra-cluster degree > 10.

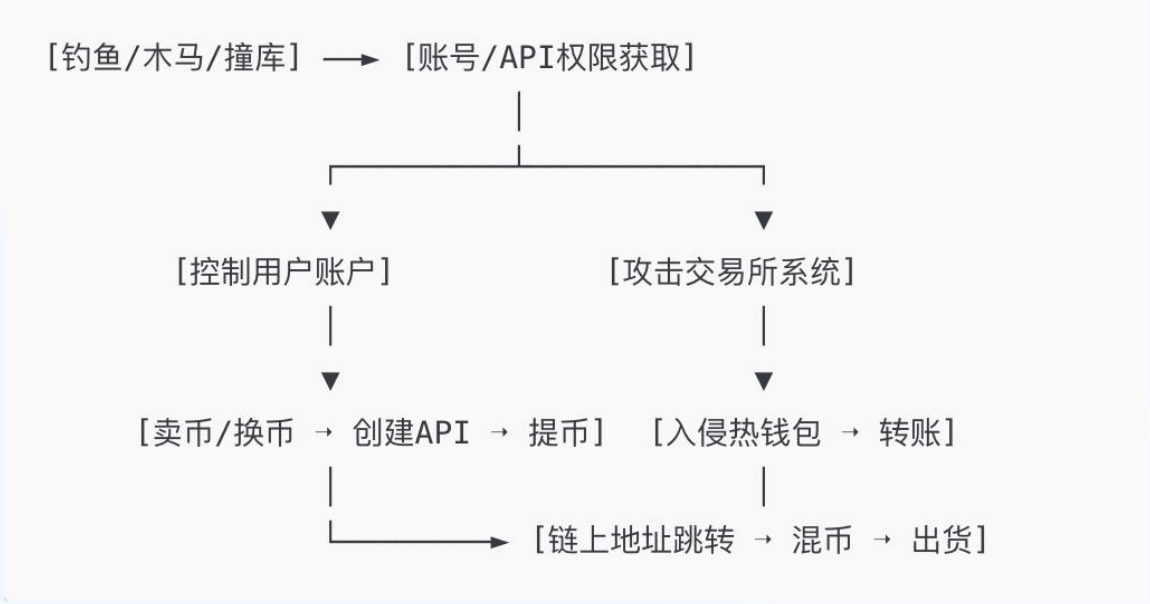
Comprehensive Risk Network Coverage
Graph algorithms construct a pan-exchange network view of users and accounts, enabling continuous monitoring of all potential risks, from single-account anomalies to complex multi-account coordinated fraud.

Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Transaction Risk Scenario Analysis

Hacker Attack Sequence Mapping



Triggering Risk Control Signals

- **Unusual Login Location**: Large IP geolocation changes within a short time frame or logins from suspicious regions
- **Device Fingerprint Anomalies**: Logins from previously unseen devices or environments
- **Rapid Operation Sequence**: Login ➞ API Key Creation ➞ Coin Sale ➞ Withdrawal — executed in quick succession
- **Abnormal Withdrawal Address**: Withdrawal to addresses never used before or addresses shared with other known malicious accounts
- **Withdrawal Amount Near Limit**: Withdrawal amounts close to system limits, often split into batches to evade detection rules

Challenges Faced by Exchanges

- **Multi-Hop Obfuscation: Hackers** use multi-hop wallet transfers to evade detection.
- **Aggregation Wallet Identification Difficulties**: Withdrawals from different users may converge at a single aggregation wallet, making attribution and tracking difficult.
- **Real-time Processing Challenges:** Traditional rules cannot respond to high-speed trading and instant operations.

Graph-Powered Solutions for Digital Currency Exchanges

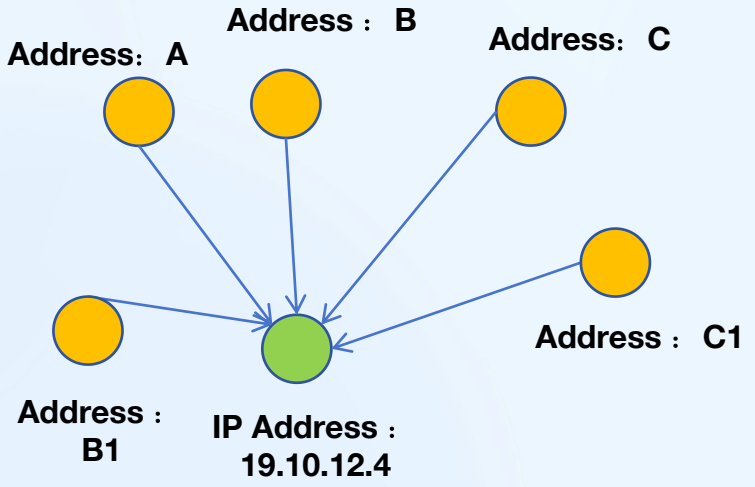
PUSHI AI

Transaction Risk Control: Real-Time Detection via Graph-Based Rules

We build a risk control graph using phone numbers, IPs, devices, wallet addresses, identities, and behavioral patterns.

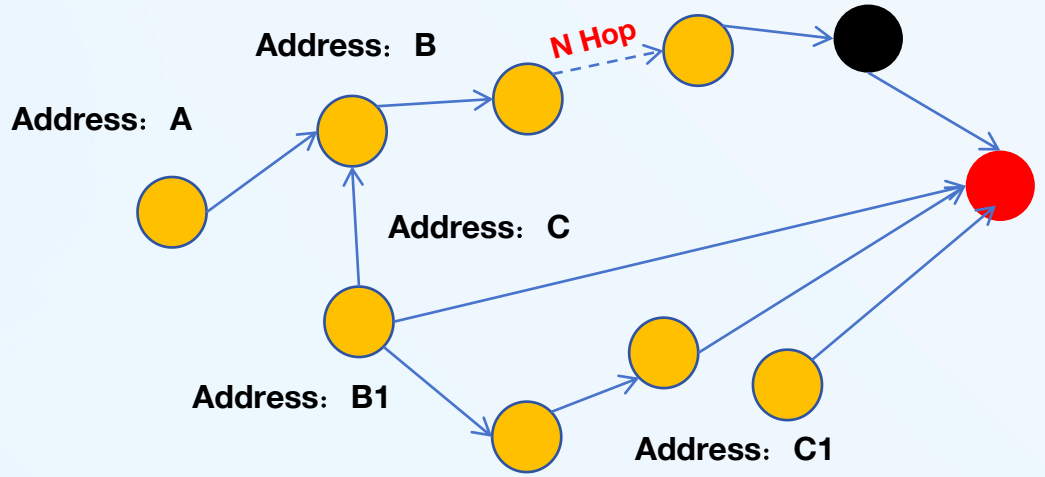
By abstracting complex address and account relationships into simplified behavioral topologies, we apply graph-based rules for real-time detection of malicious activities.

Login Behavior Pattern



Multiple accounts accessed from the same IP within 5 minutes

Fund Flow Behavior Pattern



Funds from multiple wallets converge on a single illicit receiving address via different transfer paths

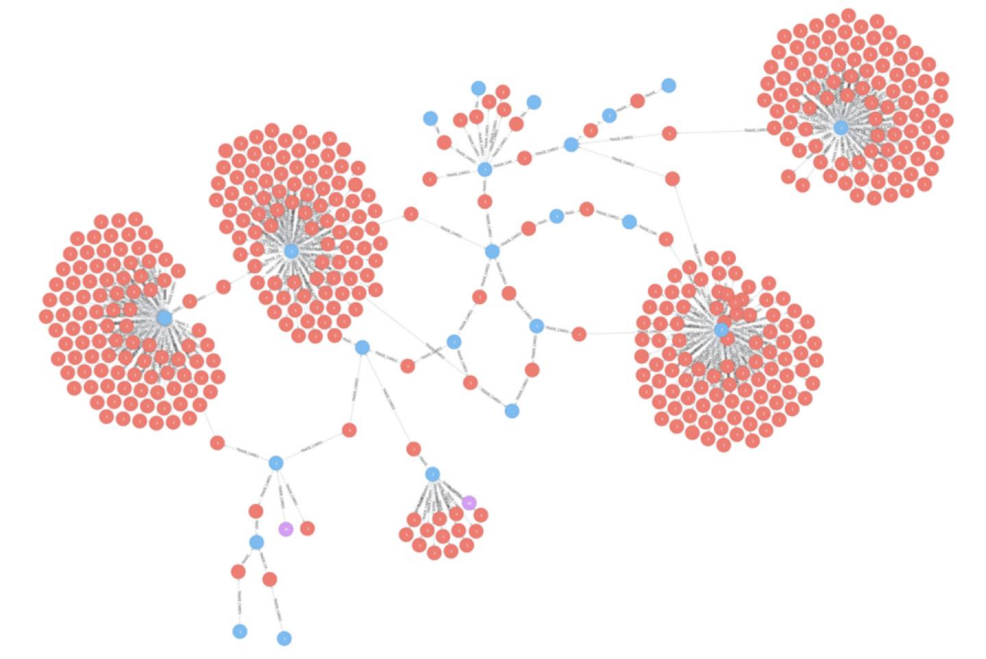
Graph-Powered Solutions for Digital Currency Exchanges

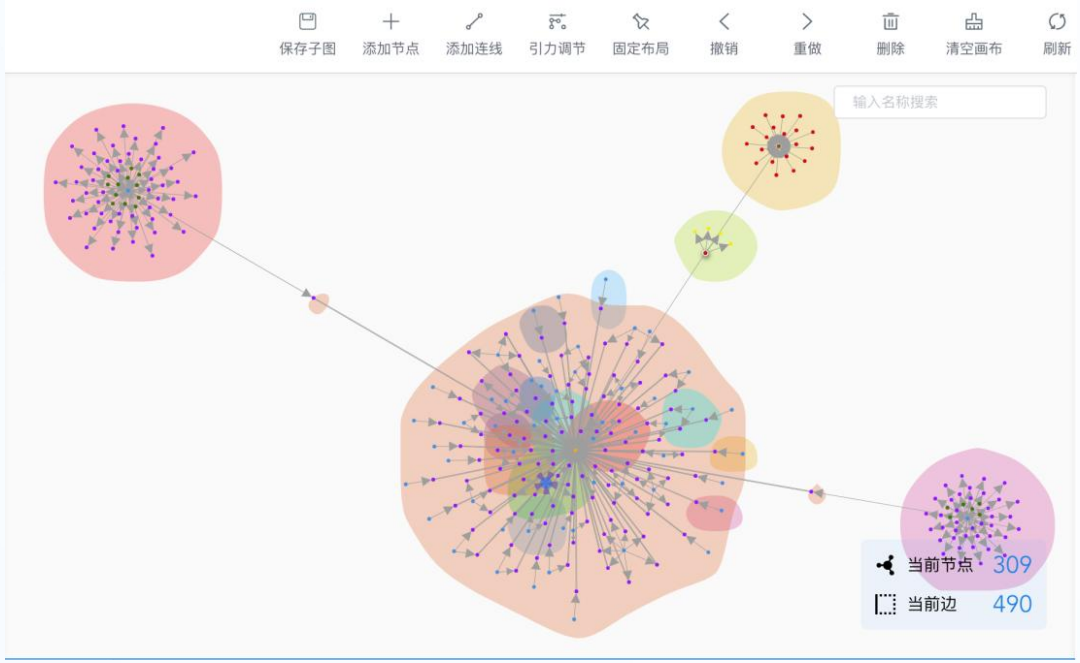
PUSHI AI

Transaction Risk Control: Identifying High-Risk Communities Using Graph Computing

Construct a graph with key entities — accounts, addresses, IPs, devices, and transactions — and apply the Louvain algorithm to detect tightly-knit high-risk communities.

By propagating labels from known malicious entities, risk signals diffuse through strong connections, uncovering hidden exposure in transaction networks.

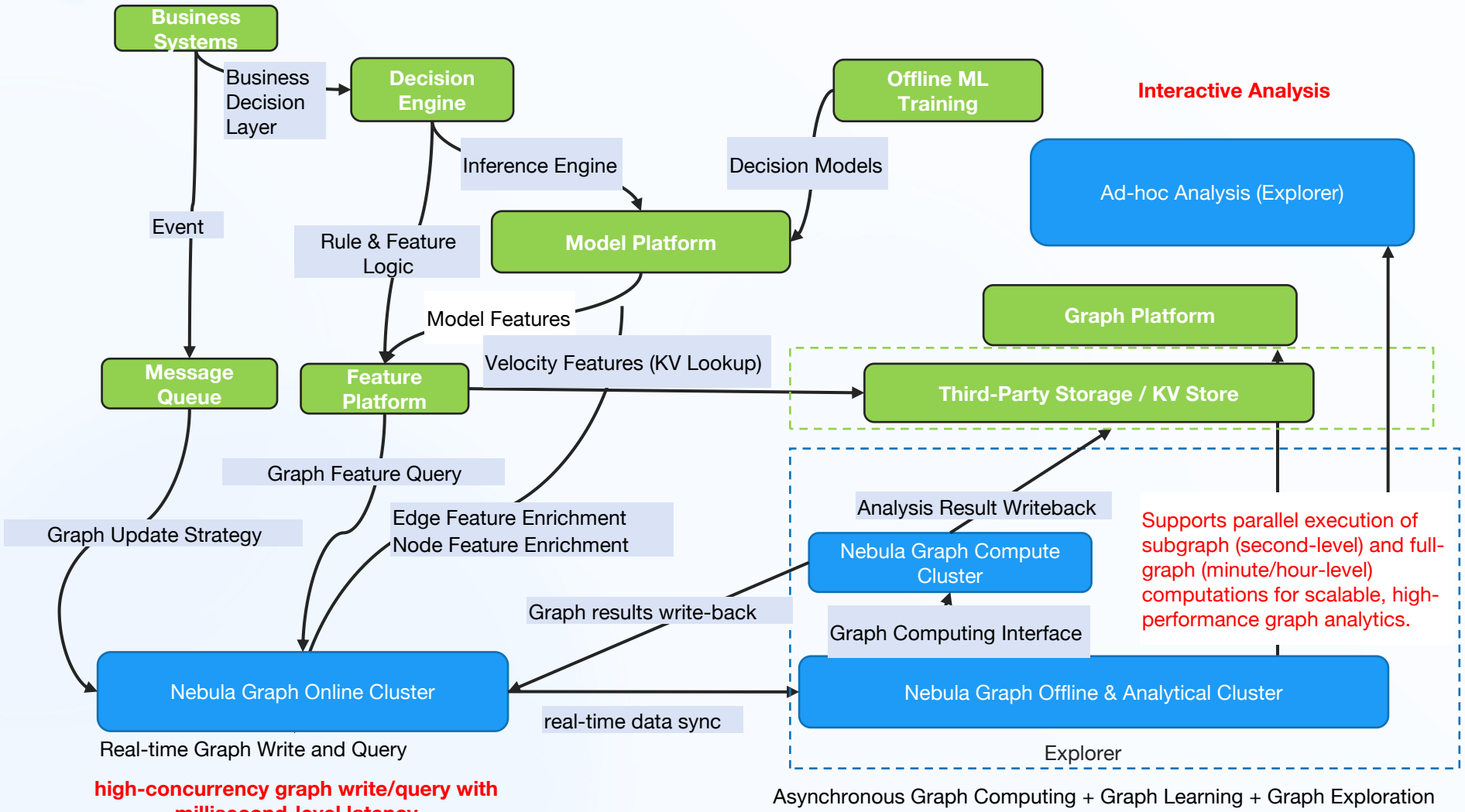




Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Hybrid Architecture for Real-Time, Near Real-Time, and Offline Processing Scenarios



1.Two dedicated clusters support both real-time operations and offline analytics.

2.The online cluster handles high-concurrency writes and low-latency queries.

3.Data is synchronized in near real time from the online to the offline cluster.

4.The offline cluster performs asynchronous and complex graph computations, with results written back to the online cluster.

5.The offline cluster supports both subgraph and full-graph analytics.

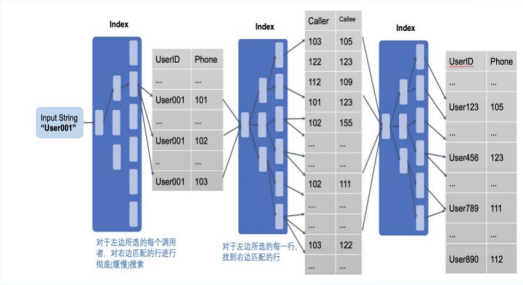
6.Analyticd cluster enables graph computing and can be dynamically scaled during off-peak hours for full-graph workloads.

Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Real-Time Risk Control Powered by Graph Technology

Graph Databases Enable Complex Correlation Metrics for Real-Time Risk Control



Graph Metric Computation Using Relational Databases

Graph Metric Computation Using Graph Databases

- **Fast Multi-Hop Risk Association Detection**
  Instantly query users' multi-hop relationships with denylists or suspicious clusters to accurately identify high-risk addresses.。
- **Enhanced Rule & Model Effectiveness**
  Graph features can be directly used in rule engines and ML models to improve precision and interception accuracy.
- **Low Latency & High Concurrency for Real-Time Decisions**
  Millisecond-level complex relationship queries support high-throughput scenarios like real-time reward distribution.

Metric	Relational DB Execution Time (s)	Graph DB Execution Time (s)	Records Return
2	0.0016	0.001	~250
3	3.0267	0.0168	~11000
4	154.3505	0.1359	~60000
5	未完成	0.2132	~80000

Metric	Concurrency	QPS	Average Latency (ms)
1	500	124081.01	1.4
2	500	143352.06	1.27
3	500	6345.76	47
4	500	227.24	126

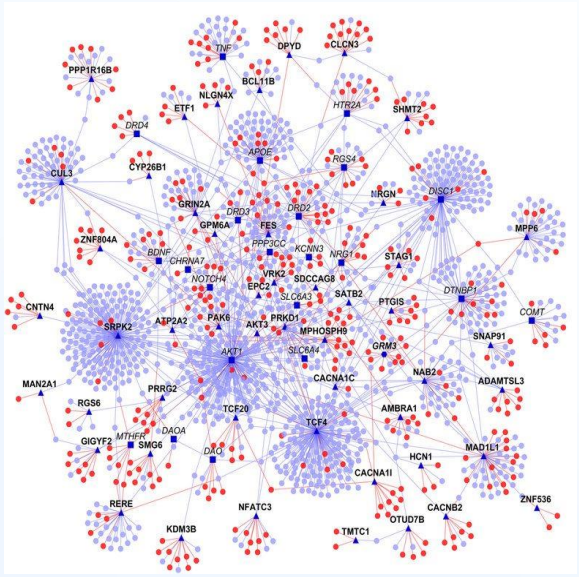
Dataset: Hundred-Million-Scale Anti-Fraud Test Data

Graph-Powered Solutions for Digital Currency Exchanges

PUSHI AI

Dynamic Strategy Generation Based on Graph Features

Leverage richer data and machine learning approaches to uncover new relational patterns, enrich graph-based features, and improve the effectiveness of fraud detection strategies.



Richer Features, Higher Accuracy

- Augment traditional statistical metrics with relational (graph-based) features.
- Graph features reflect real-world connections, making them more resistant to manipulation.

Explainable

- Break away from black-box AI — graph-based features are inherently interpretable, supporting transparent decision-making.

Graph Feature Extraction

Enriching Node and Edge Features

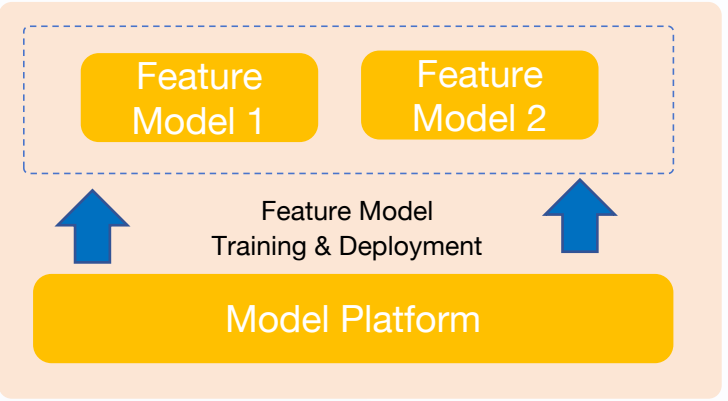
Feature Enrichment

- Add similarity-based relationships in addition to direct entity connections.
- Supplement graph nodes with tag or label features.

Enhanced Query & Computation Capabilities

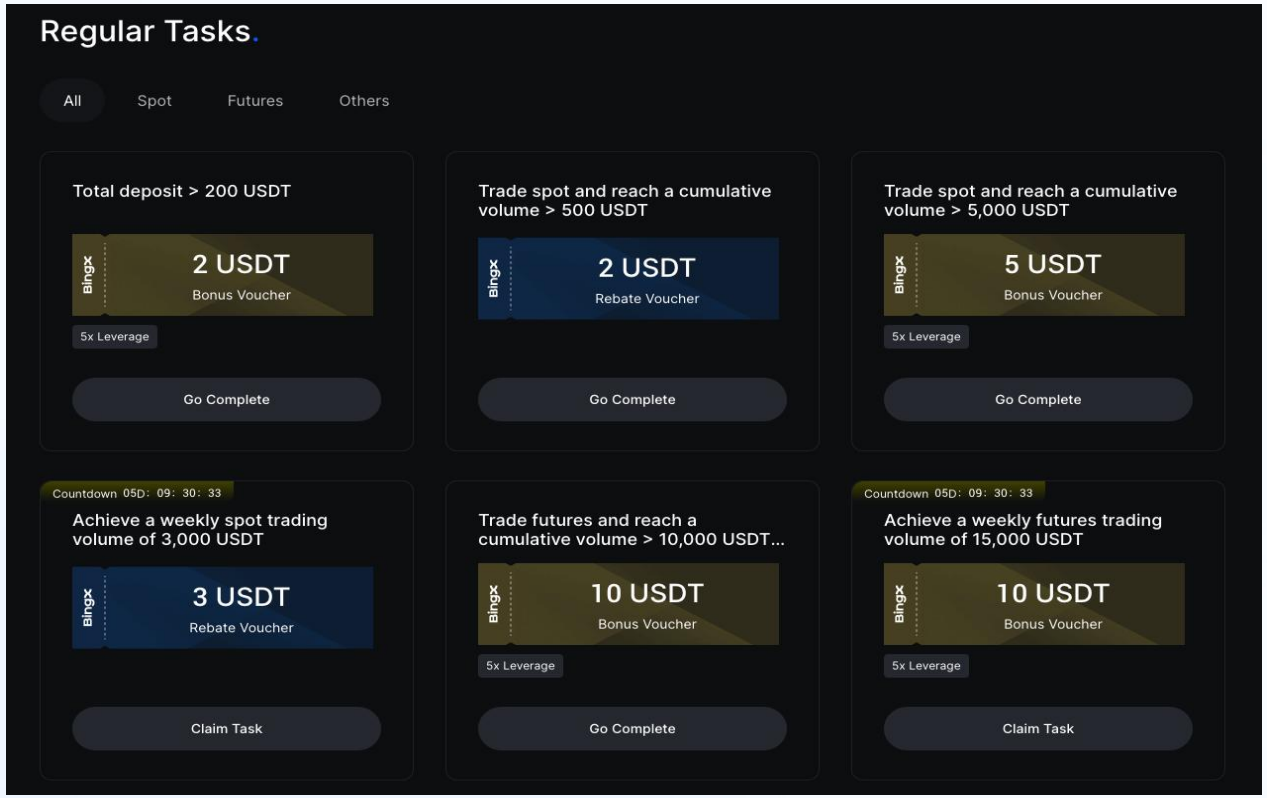
- Leverage expanded relational dimensions to improve the depth and accuracy of graph queries and computations.

Feature Platform



Successful Use Cases of Graph Databases in the Web3 Industry

Case 1：Crypyto Exchange, CSI Platform Powered by a 10-Billion-Scale Graph



Business Background

•Stricter Global Regulations
Regulatory bodies like the SEC and ESMA are tightening controls on crypto and DeFi.

•Enhanced AML/KYC Requirements
Exchanges enforce stronger AML and KYC to block illicit funds.

•High Trading Volume
With billions traded daily, detecting suspicious transactions is crucial.

Business Challenges

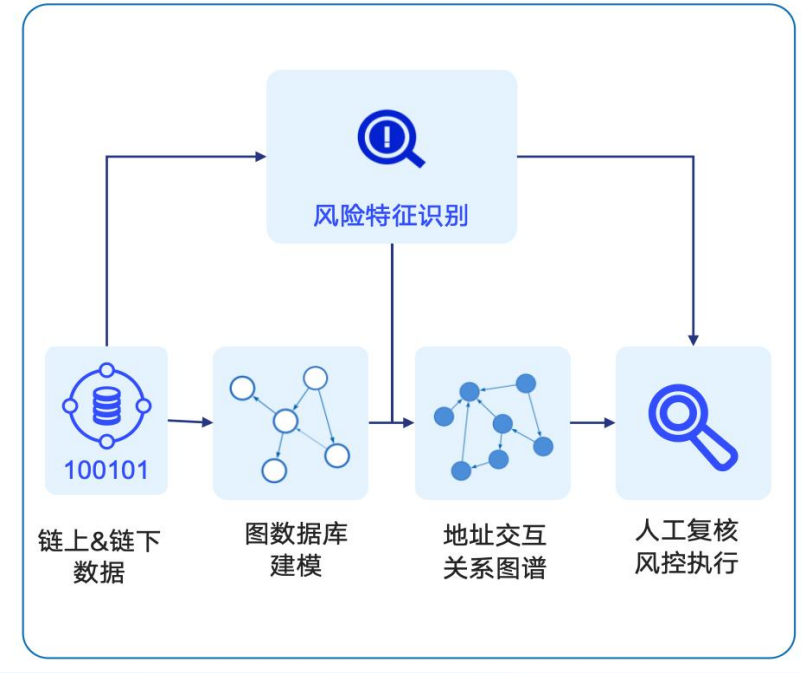
•Compliance Pressure
Stricter AML and KYC requirements demand more efficient identity verification and fund monitoring to prevent illicit flows.

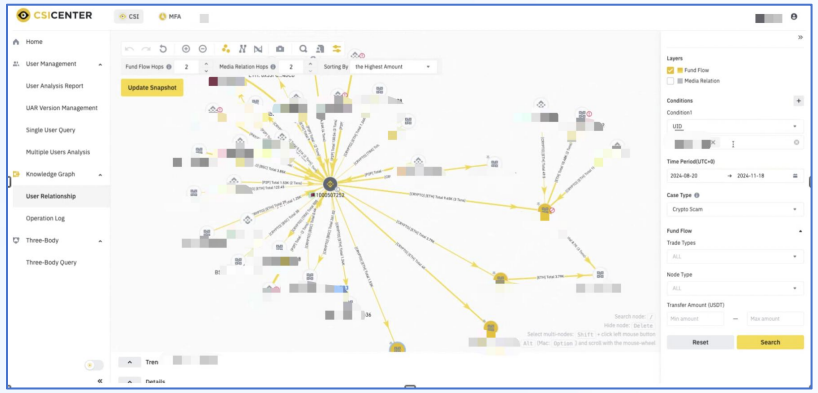
•Massive Data Volume
With daily trading volumes in the tens of billions, identifying suspicious activities like money laundering or market manipulation has become a major technical challenge.

Successful Use Cases of Graph Databases in the Web3 Industry

PUSHI AI

Case1：Crypyto Exchange, CSI Platform Built on a 10-Billion-Scale Graph





Graph-Based Risk Control Solution

- **Massive-Scale Risk & Compliance Graph,** built using off-chain data, on-chain data, user profiles, and transaction records — totaling 4 billion nodes and 40 billion edges.
- **Real-Time Risk Detection,** graph metrics are queried in real time at user registration to detect potential bonus abuse or suspicious users
- **Offline Graph Analytics,** full-scale analysis of on-chain and off-chain data to uncover complex suspicious transaction patterns.
- **Visual Investigation Tools**
  Analysts can perform manual reviews using graph visualizations combined with high-risk watchlists.

Value Proposition

•**Real-Time and Accurate Risk Monitoring i**nstantly detects suspicious transactions and abnormal fund flows from massive data streams, improving the speed and accuracy of fraud detection.

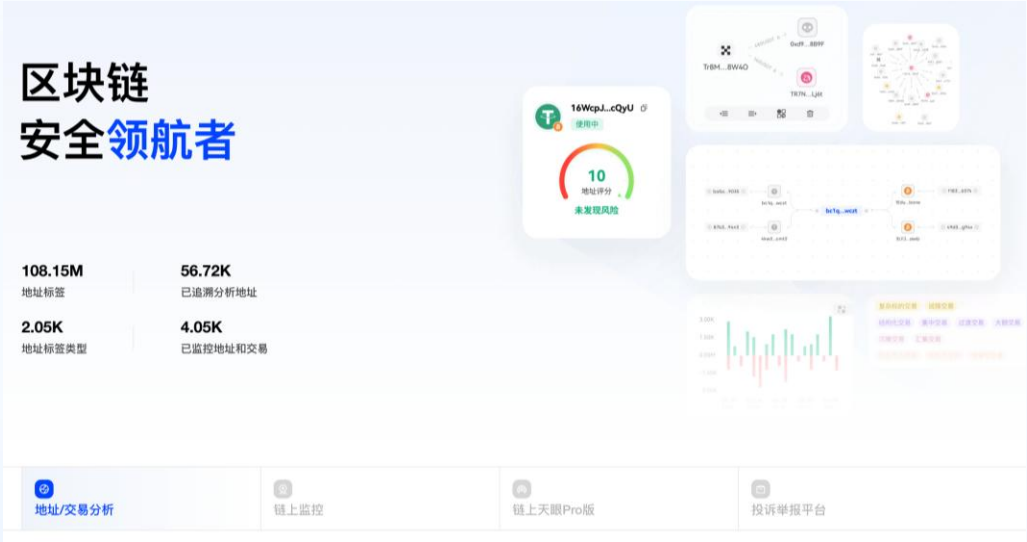
•**Efficient Cross-Account Correlation** uncovers complex relationships between accounts to enhance cross-platform risk identification and prevent illicit networks.

•**Improved Compliance Transparency & Auditability**
Automates compliance data aggregation and reporting, ensuring clear, explainable, and traceable compliance for audits and regulatory reviews.

Successful Use Cases of Graph Databases in the Web3 Industry

Case2：Providing Personalized On-Chain Data Services for Users

In the crypto world, whoever can efficiently collect and harness massive on-chain data gains a first-mover advantage at the transaction level.



Business Background

- customer has labeled over **3 billion addresses** and streams real-time on-chain data, enabling native transaction analysis and address tracing.
- With rich data on address tags, risk profiles, and transactions, it delivers services like **address scoring**, **profiling**, and **graph-based analysis**.

Business Challenges

•How to extract valuable insights from massive, entangled on-chain data at scale.

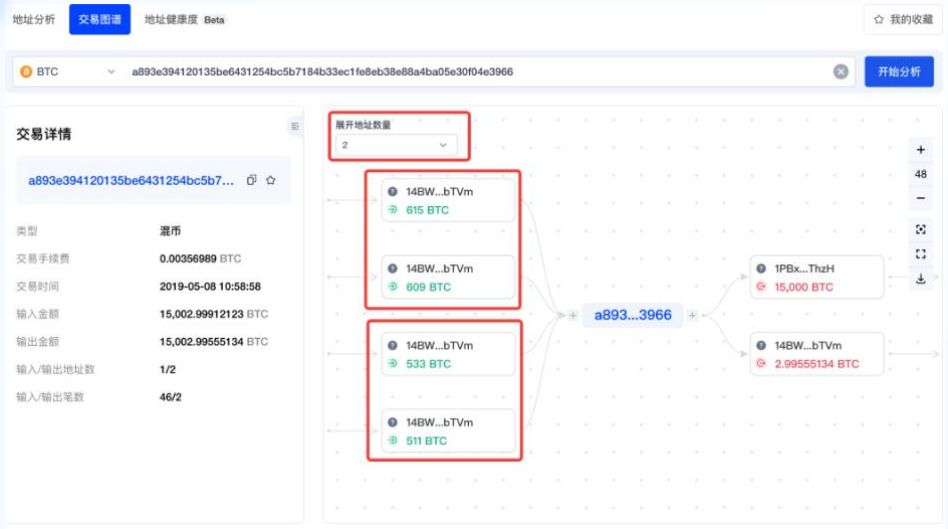
•Complex multi-chain transactions and multi-hop anonymous paths make traditional tracking methods ineffective.

•Deep path exploration suffers from high latency and slow response, making it difficult to reach the final destination address.

Successful Use Cases of Graph Databases in the Web3 Industry

PUSHI AI

Case2: Providing Personalized On-Chain Data Services for Users



All-in-One Graph Platform Solution

- **Graph Modeling**: Unified modeling of heterogeneous entities such as addresses, transactions, and contracts
- **Real-Time Graph Computation**: Path analysis via graph database in real time
- **Graph Algorithms**: Fast detection of fund aggregation using path search and community detection
- **Graph Visualization**: Intuitive tracing of transaction paths
- **Graph + AI**: Training label models with graph features to enable automated iteration

Business Value

- **Broader Asset Coverage**: By leveraging graph databases' strength in relationship storage, the address/transaction analysis capabilities of the on-chain monitoring system have expanded beyond Bitcoin, Ethereum, and OKC to cover more asset types.
- **Address Risk Profiling**: Graph rules, metrics, and algorithms help uncover suspicious transaction patterns to assess address "health."
- **Deep On-Chain Graph Analysis**: Powerful graph computations enable multi-hop transfer tracing and deeper investigation into complex transaction paths.

contents

PART ONE About Us

PART TWO Knowledge Graph-Based Web3 Risk Control Strategies

PART THREE Product Overview

PUSHI Product

Empower Enterprises to Deliver Measurable Business Outcomes

PUSHI Graph Intelligence Platform - An end-to-end AI system delivering measurable business outcomes through integrated technologies

Data Analytics Reinvention

- **Multi-Source Integration**: Seamlessly combines structured and unstructured data from diverse sources.
- **Automated Knowledge Graph Construction**: Rapidly extracts entities and relationships with high accuracy.
- **Advanced Graph Analytics**: Integrates graph algorithms (e.g., community detection, path analysis) and rule engines to uncover hidden data connections.
- **Intuitive Visualization**: Translates complex relationships into interactive network diagrams using graphical interfaces and visualization tools.

Data Analysis

- **Zero-Code Platform**: Enables non-technical users to perform sophisticated analyses through drag-and-drop modeling.
- **User-Friendly Interface**: Simplifies complex workflows while maintaining enterprise-grade analytical depth.

Flexible and Extensible Architecture

- **Modular Design**: Supports integration with machine learning, large language models (LLMs), and privacy-preserving technologies.
- **Cross-Industry Adaptability**: Preconfigured solutions for finance (risk control, AML), public security (criminal investigation), healthcare (patient journey mapping), and IT operations.

Enterprise AI Agent Capabilities

- **Graph-LLM Synergy**: Combines graph intelligence with large language models for contextual decision-making.
- **Knowledge Base Integration**: Manages documents, data tables, and graph knowledge to empower AI agents with multi-modal insights.

PUSHI Product

Empower Enterprises to Deliver Measurable Business Outcomes

PUSHI Graph Intelligence Platform - An end-to-end AI system delivering measurable business outcomes through integrated technologies

Knowledge Graph Construction

Multi-Source Integration for Foundational Knowledge Graph Construction

Rule-based Reasoning, Predicate Filtering, Temporal Filtering

Dynamic Business Scenario Graph Generation

Knowledge Graph Inference

Graph Rule Matching

Graph Metrics Calculation

Graph Algorithms

Graph Machine Learning

Multi-modal graph analysis detects and ranks high-risk entities/groups in business networks.

Visual Analytics

Knowledge Graph Visual Analytics

Conduct expert validation of graph reasoning outputs and utilize interactive visual analytics to uncover hidden patterns.

Clue Analysis

Case Investigation

Expert-validated subgraphs are stored in the Knowledge Graph Case Repository for compliance audits, report generation, collaborative analysis, and case resolution.

Intelligent Agent Development

Application Deployment

An intelligent agent factory enabling enterprise acceleration through diverse agent types, multi-domain knowledge orchestration, and agile deployment capabilities.

Next-Generation AI Application Development Platform

PUSHI AI

Lowering Barriers to AI Adoption · Empowering Enterprise Intelligent Transformation

Core Technologies:

Graph Computing:

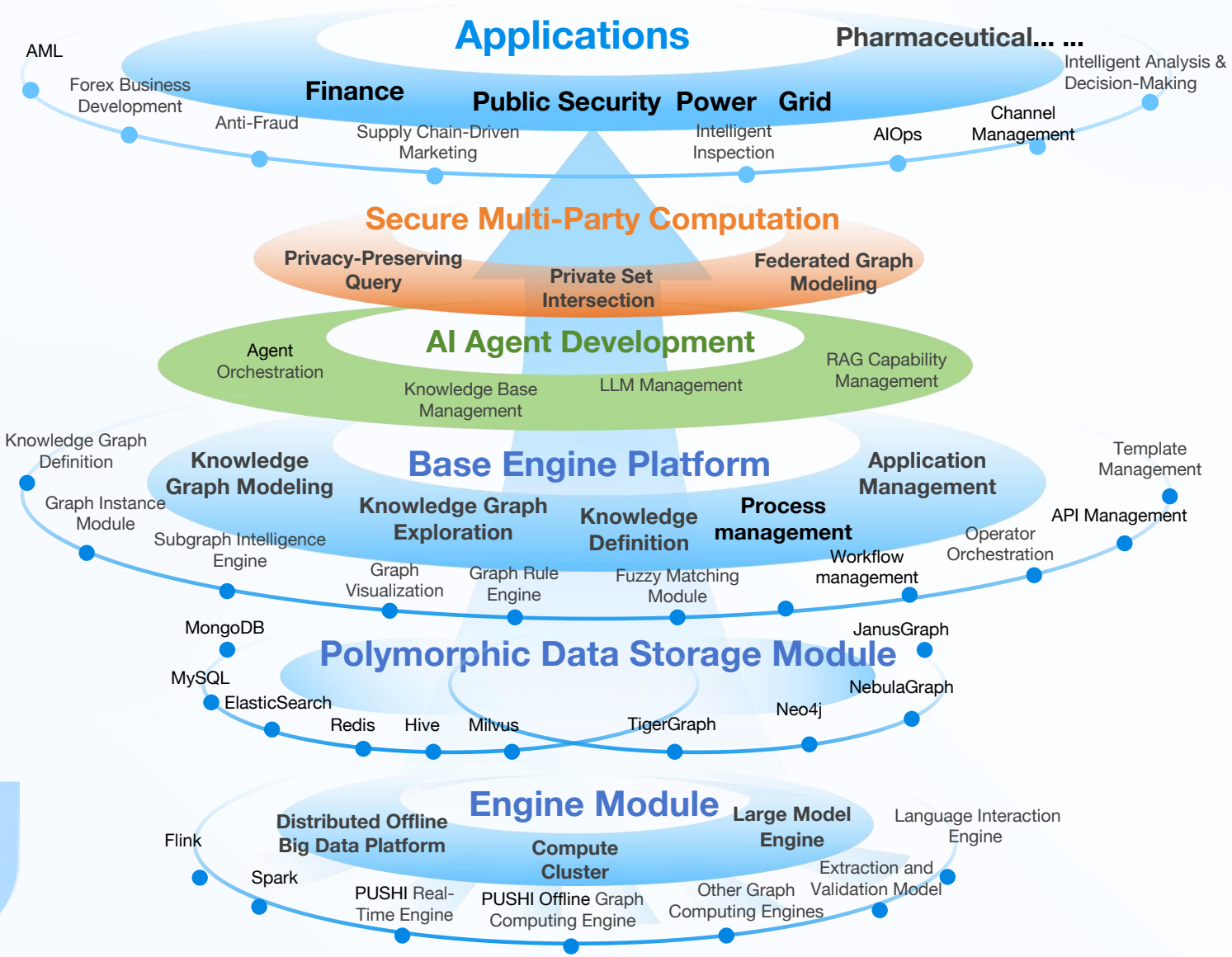
• Models complex relationships via vertices and edges to uncover hidden patterns in networks.
•Converts unstructured data into graph structures, transcending traditional flat-data analysis.

Intelligent Agents: Amplifies knowledge value and delivers personalized services through AI-driven automation.

Privacy-Preserving Computation: Expands data collaboration securely across organizations, enabling trusted multi-party synergy (1+1>2 effect).

Ultimate Vision

AI Agent-Driven Decision-Making Production Line



Core Capabilities

1.Multi-Source Data Integration

- **Data Connectivity**: Supports mainstream relational databases (MySQL/Hadoop), APIs, and streaming data ingestion
- **Automated Schema Mapping**: Prebuilt templates reduce graph transformation costs

2. Zero-Code Graph Building

- **Visual Ontology Modeling**: Drag-and-drop ontology modeling
- **Intelligent Entity Alignment**: Cross-source mapping accuracy (industry-leading)
- **Dynamic Temporal Graphing**: Time-aware relationship version control

3. Interactive Intelligent Analytics Hub

- **Trillion-Scale Relationship Analysis**: Multi-dimensional querying and subgraph annotation
- **Rule Engine Clusters**: NLP-to-graph query conversion lowers technical barriers to use
- **Pre-Configured Algorithms**: 30+ graph models for community detection, risk propagation, etc.

4. Metric-Driven Decision Empowerment

- **Visual Metric Development**: Custom calculations and deployment of rule-based indicators
- **Adaptive Threshold Alerts**: Auto-generated monitoring metrics from graph features
- **Model Servitization:** Export metrics to BI/risk systems via APIs

5. Enterprise-Grade Process Orchestration

- **Workflow Designer**: 200+ node pipeline visualization and automation
- **AI-Optimized Scheduling**: Dynamic resource allocation boosts task efficiency
- **Full Lifecycle Monitoring**: 100% task status traceability

6. Domestic Innovation Ecosystem Integration

- **Localized Compatibility**: Certified for Kunpeng/Kylin/Muxi/Dameng infrastructure
- **Secure Data Collaboration**: Federated modeling across heterogeneous databases
- **Standardization**: Recognized in CAICT's *High-Quality Digital Transformation Solutions Panorama*

普适智能
PUSHI AI

THANKS

商务咨询：marketing@pushiai.com

普适智能 WWW.PUSHI AI.COM

Apache Doris 4.0: Data Warehouse in the Era of AI

Wenqiang Li (Gabriel)

Engineer @VeloDB
Apache Doris PMC Member

Agenda

- Background
- How Apache Doris Embraces the New Era
- Why Apache Doris is the Answer
- Deep Diving into Apache Doris 4.0

Background: Legacy means Lags

Busniess Intelligence

Data is becoming ever more important and an indispensable asset

Modern Data Warehouses

More flexible architecture, capability of instant analysis and pay-as-you-go costs

New Mission in the AI Era

AI era is coming...

1990s

2022

Now

Challenges: the demand for faster and faster analysis coexists with the demand for lower and lower costs

Challenges: Requirements for Data Freshness, Real-time Analysis, Concurrent Query, AI supporting

Legacy Data Warehouse (BI-Driven)

Scalability

Expensive hardware and limited scalability.

Cost

Traditional on-premise deployment and high maintenance costs.

Latency

Limited support for real-time analytics

Worse Flexibility

Only support **internal** data source

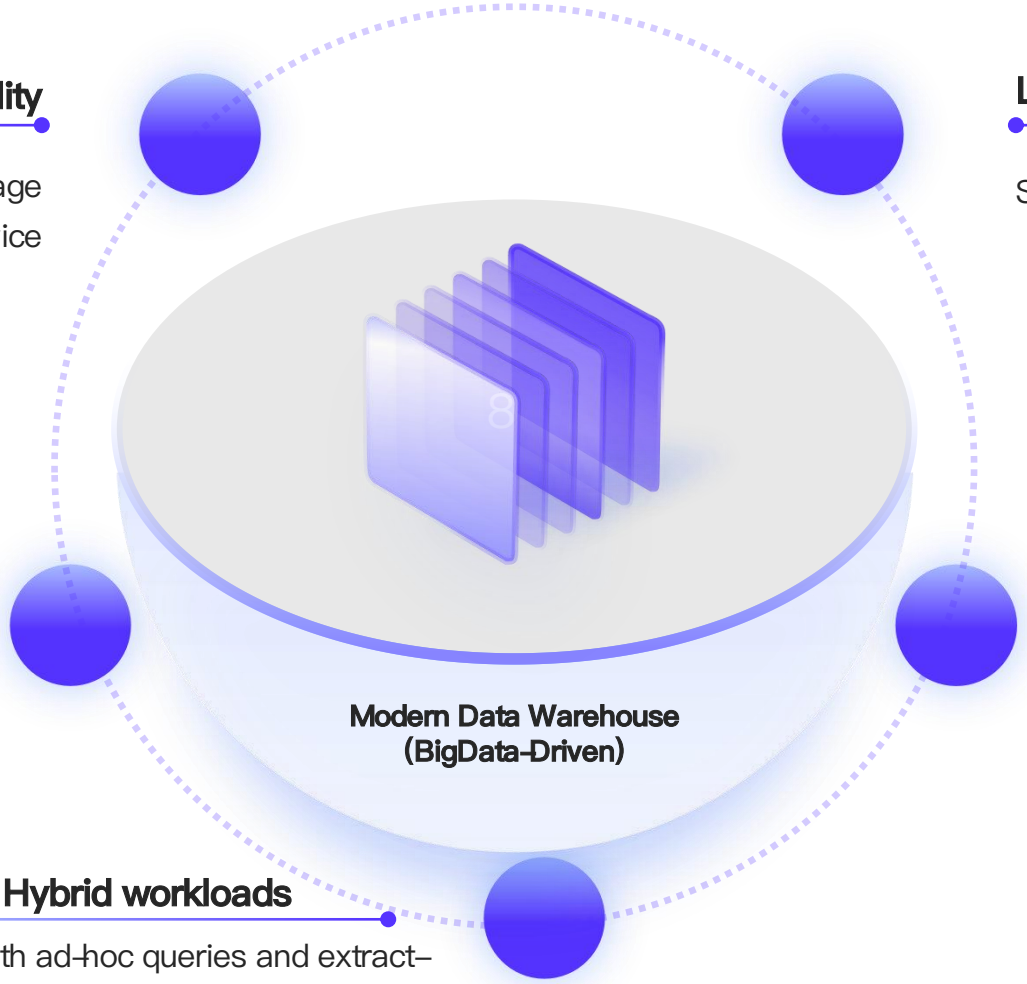
Modern Data Warehouse (BigData-Driven)

Scalability

Statelessful comptation and storage service

Low-latency

Sub-second latency



Federated Queries

Eradicate multiple data silos

High-concurrency

High-concurrent queries processing

Hybrid workloads

Both ad-hoc queries and extract–transformation-loading workloads

Data Warehouse in the Era of AI

- Both volume and quality are important
- Large amounts of data for both training and inference
- Realtime read/write and freshness requirements
- Agent-friendly
- AI as the first citizen

"Through 2027, almost all enterprises developing applications based on GenAI will invest in data platforms with vector search and retrieval-augmented generation to complement foundation models with proprietary data and content."
—From ISG Research, 2025

HowApache Doris Embraces the New Era

What's new in Apache Doris 4.0

1. Data-intensive applications and BI applications

2. Agentic-facing analytics

Use cases

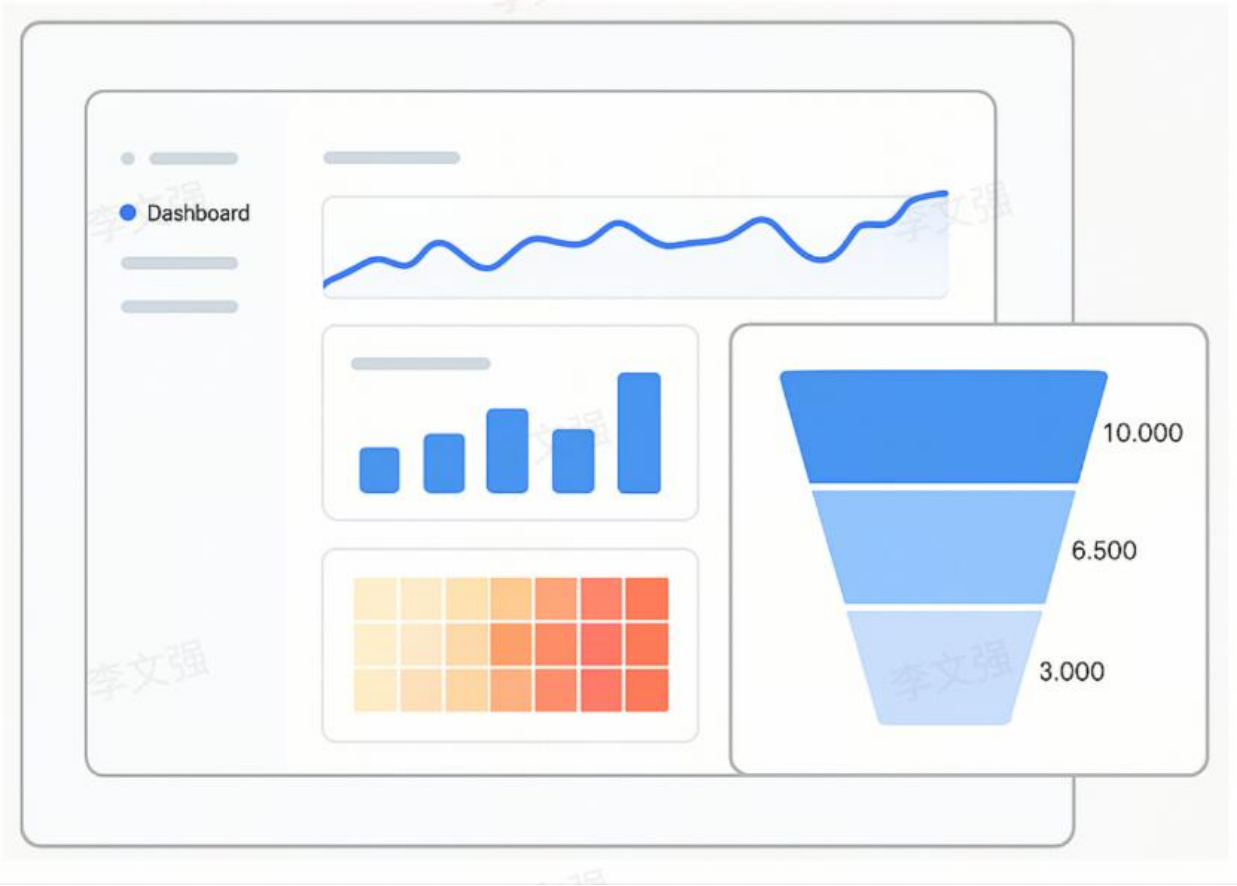
3. Federated querying

4. Powerful vector search capability

What's new in Apache Doris 4.0

Use cases

- **1. Data-intensive applications and BI applications**
- 2. Agentic-facing analytics
- 3. Federated querying
- 4. Powerful vector search capability



What's new in Apache Doris 4.0

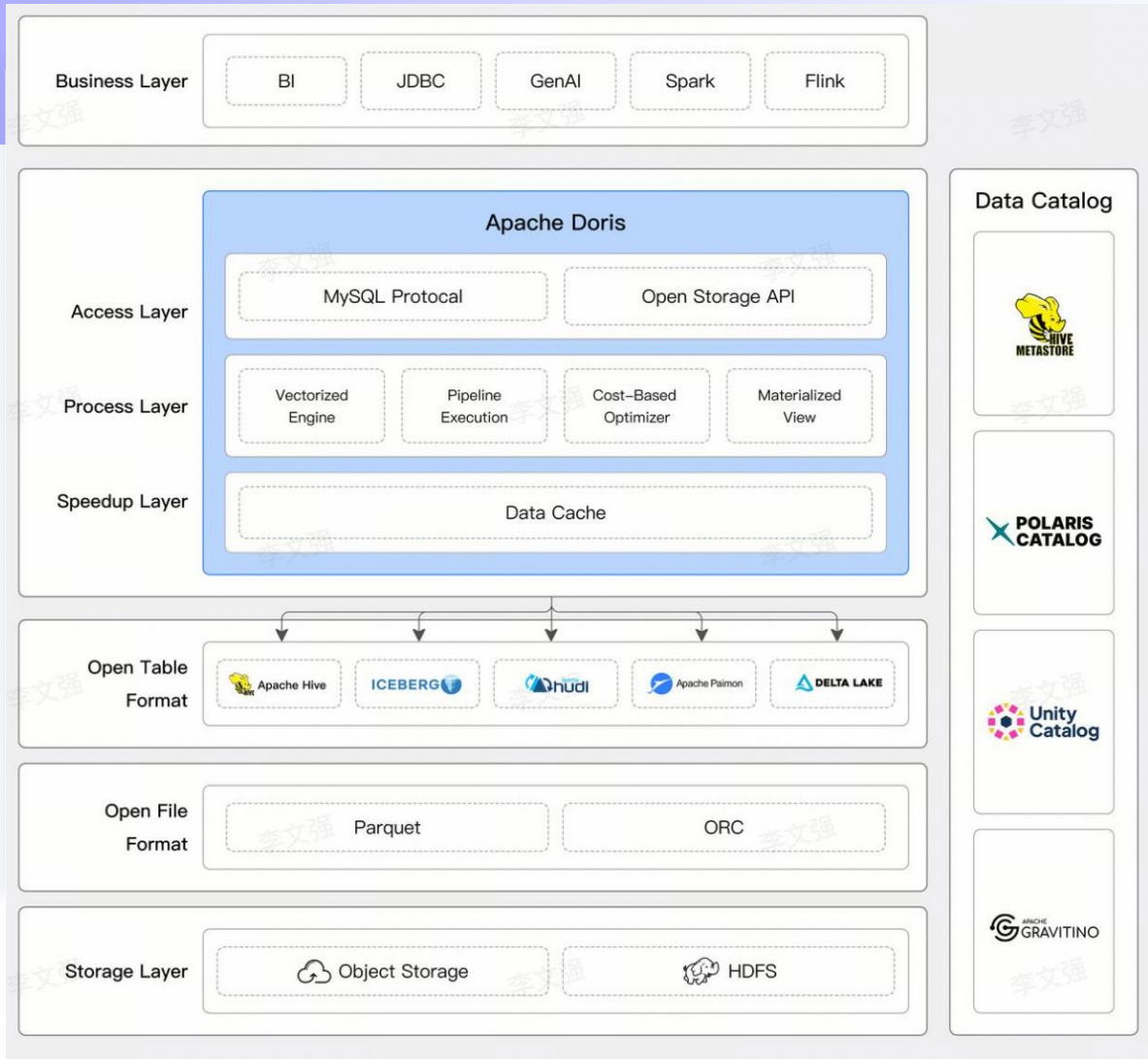
Use cases

- 1. Data-intensive applications and BI applications
- **2. Agentic-facing analytics**
- 3. Federated querying
- 4. Powerful vector search capability

What's new in Apache Doris 4.0

Use cases

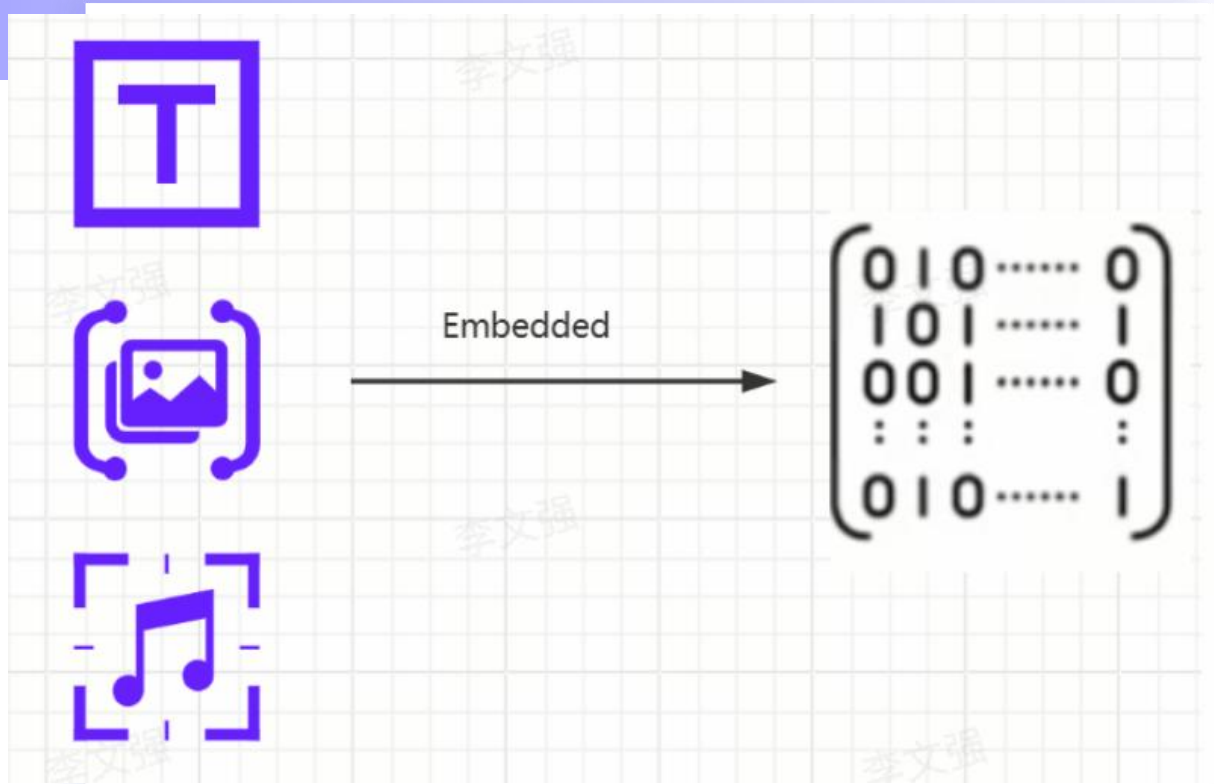
- 1. Data-intensive applications and BI applications
- 2. Agentic-facing analytics
- **3. Federated querying**
- 4. Powerful vector search capability



What's new in Apache Doris 4.0

Use cases

- 1. Data-intensive applications and BI applications
- 2. Agentic-facing analytics
- 3. Federated querying
- **4. Powerful vector search capability**



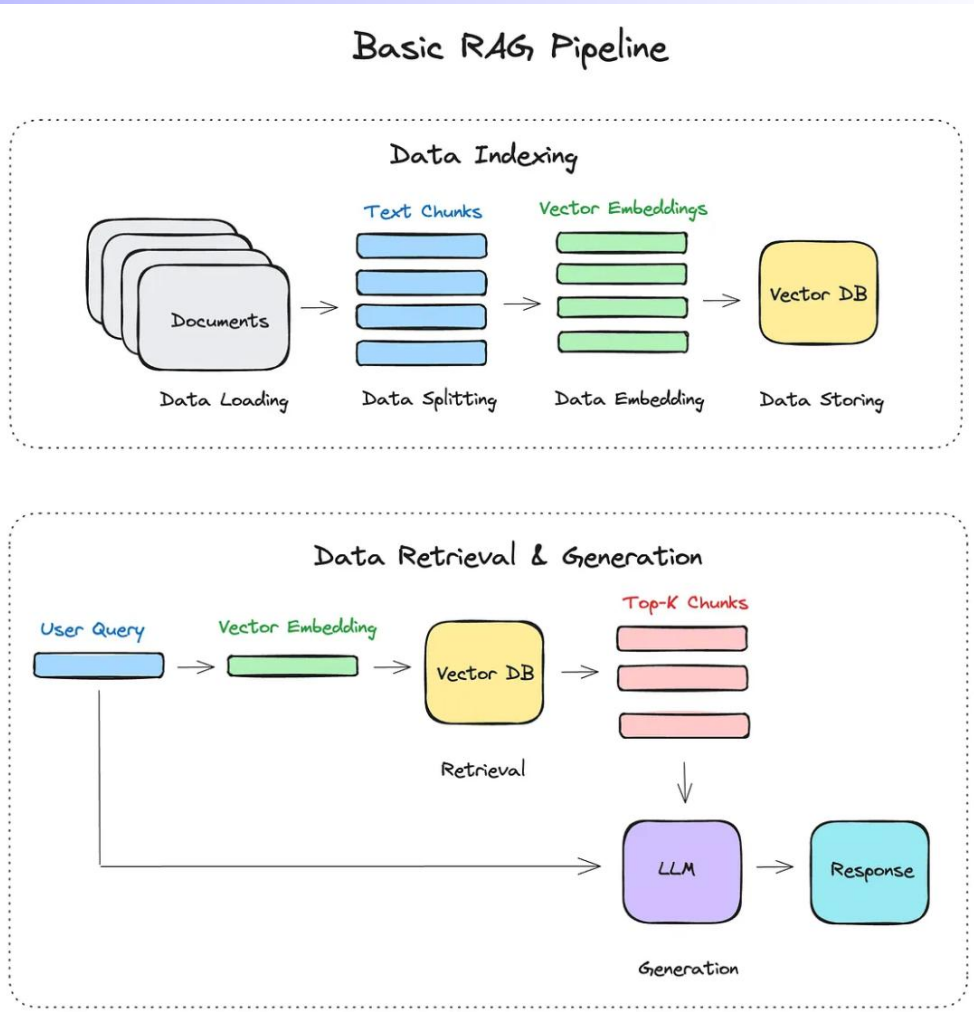
How does a database affect the performance of a model?

Challenges:

Hallucination

Solution:

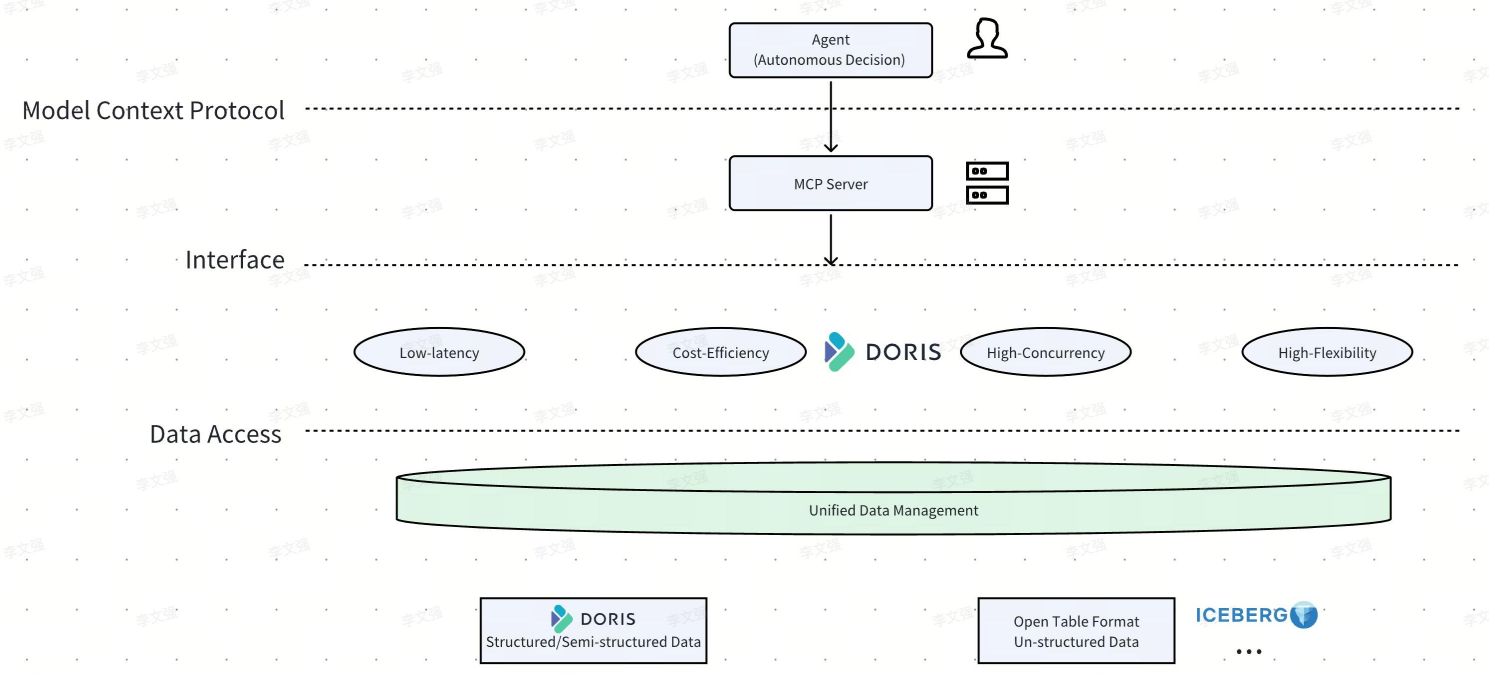
Retrieval Augmented Generation



Model Context Protocol

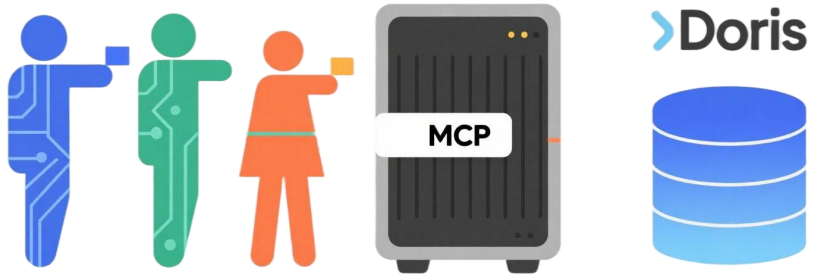
MCP like a USB-C port for AI applications. Just as USB-C provides a standardized way to connect your devices to various peripherals and accessories, MCP provides a standardized way to connect AI models to different data sources and tools.

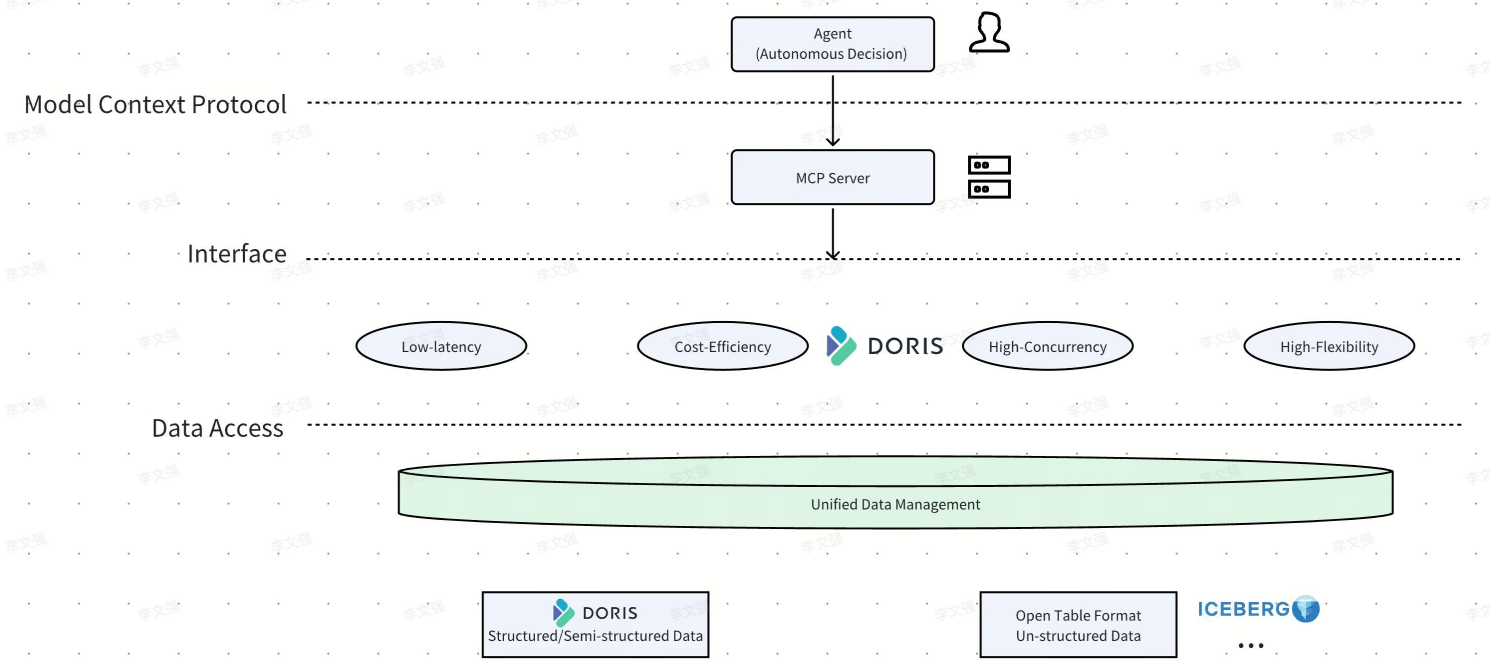
MCP: Model Context Protocol is an open protocol that standardizes how applications provide context to large language models (LLMs).



Apache Doris 4.0 Model Context Protocol

Advanced: complex decision-making task





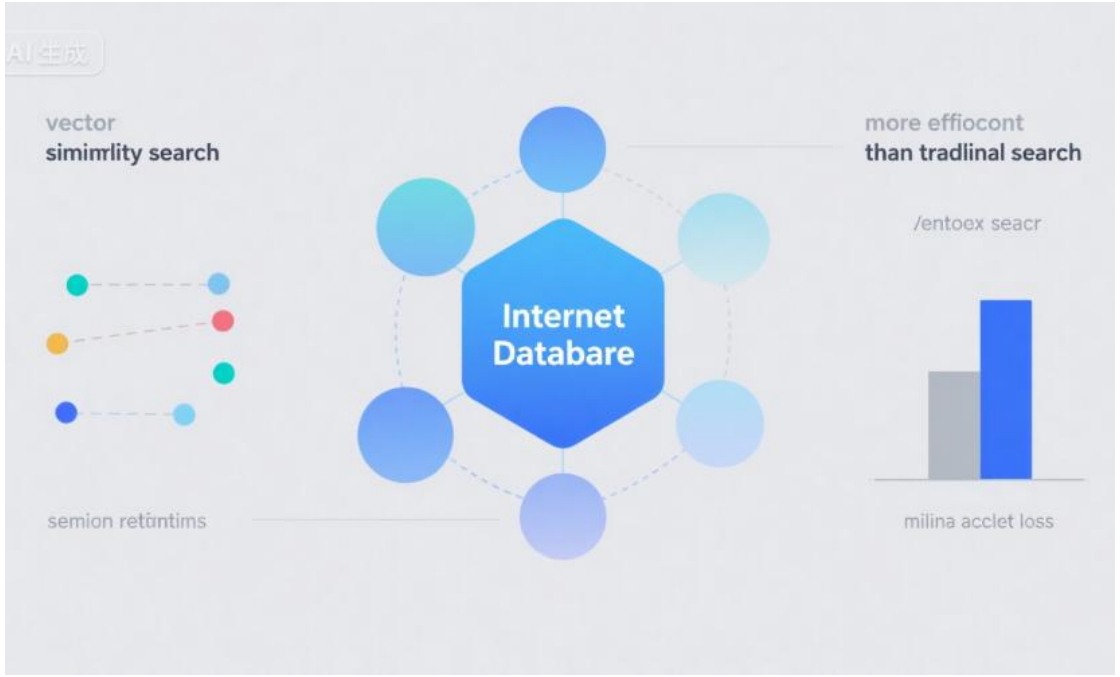
State-of-the-Art MCP Server

1. Intelligent agent orchestration
2. Revolutionary capability
3. 17 professional tools (database management, table structure analysis, query execution, performance monitoring...)
4. 8 advanced analysis tools
5. ADBC Interface (10-100 times better than traditional JDBC)

Vector Search

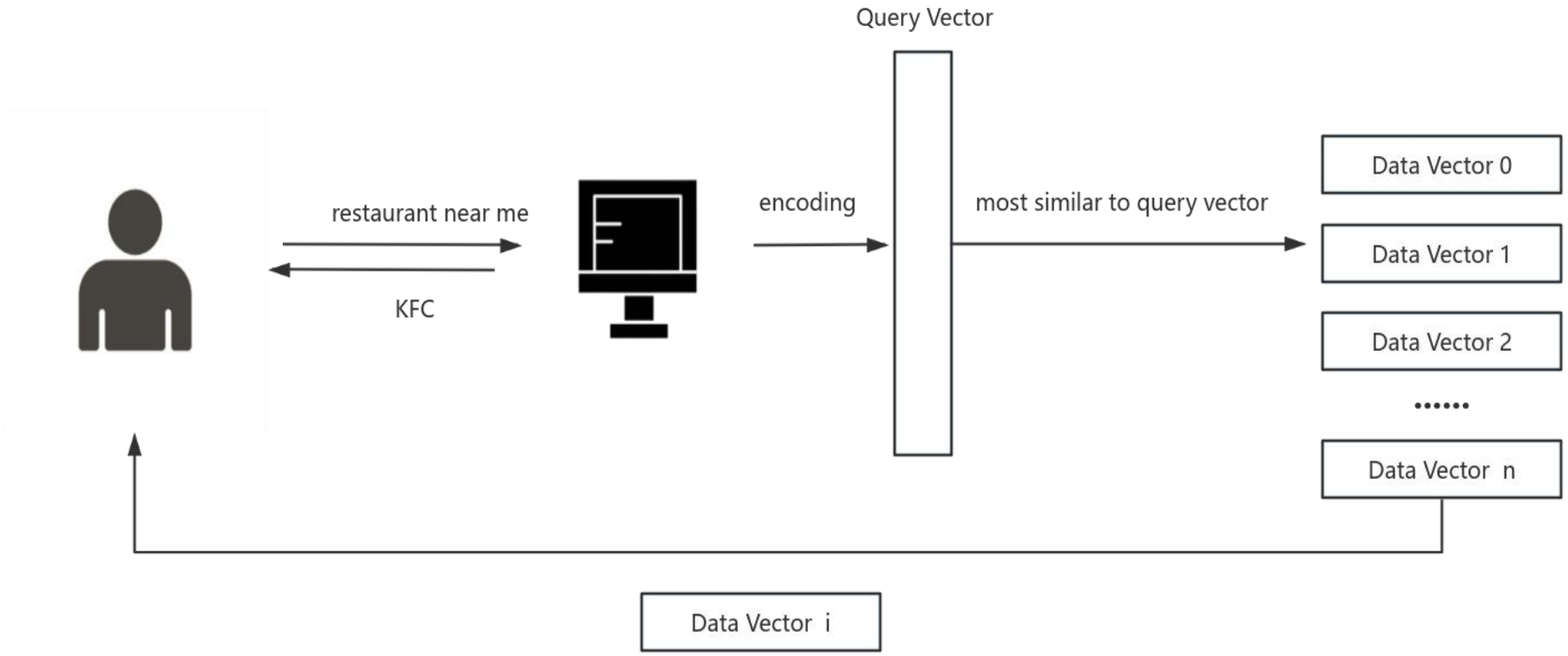
Characteristics

1. Vector similarity search
2. Semantic relationships and similarities between data points
3. More efficiency than traditional search with little accuracy loss

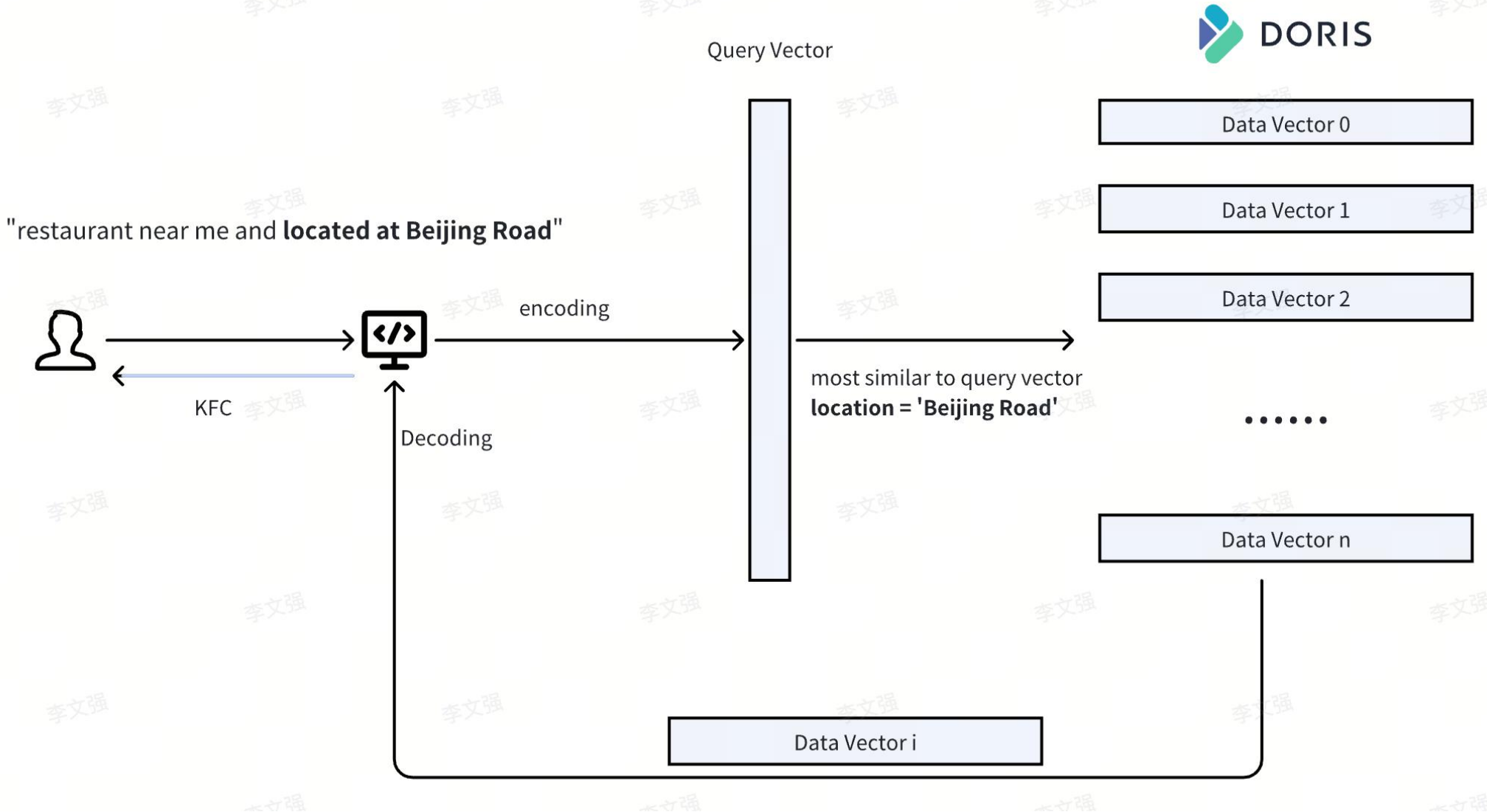


Vector search is a search technique used to find similar items or data points, typically represented as vectors, in large collections.

Vector Search



Hybrid Search

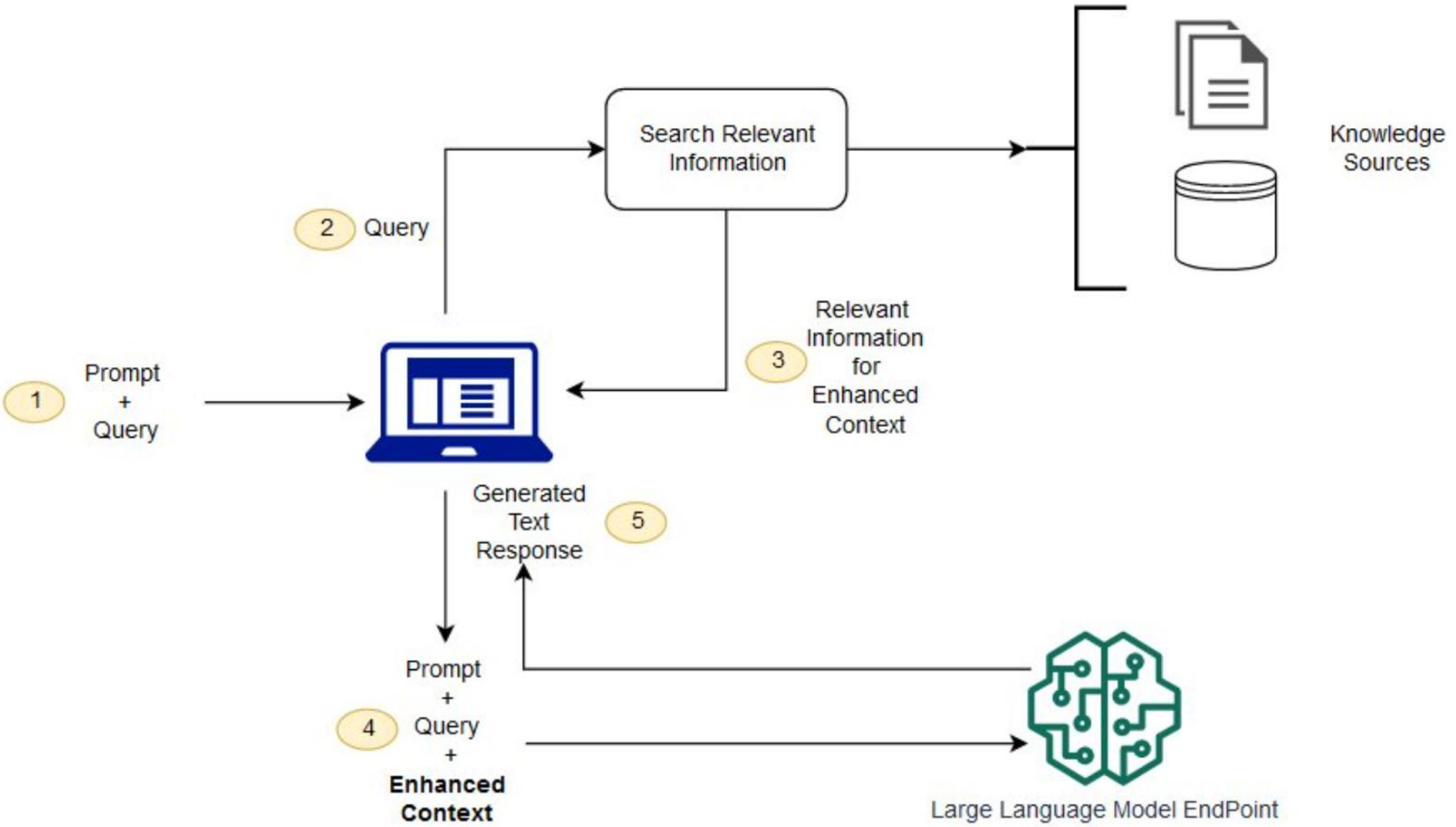


Data-related hallucinations

Large Language Model
hallucination caused by

Data-related hallucinations
Solution: Retrieval Augmented Generation

Retrieval-Augmented Generation



Retrieval-Augmented Generation

One-Time Retrieval

Used in some simple cases like translation tasks.

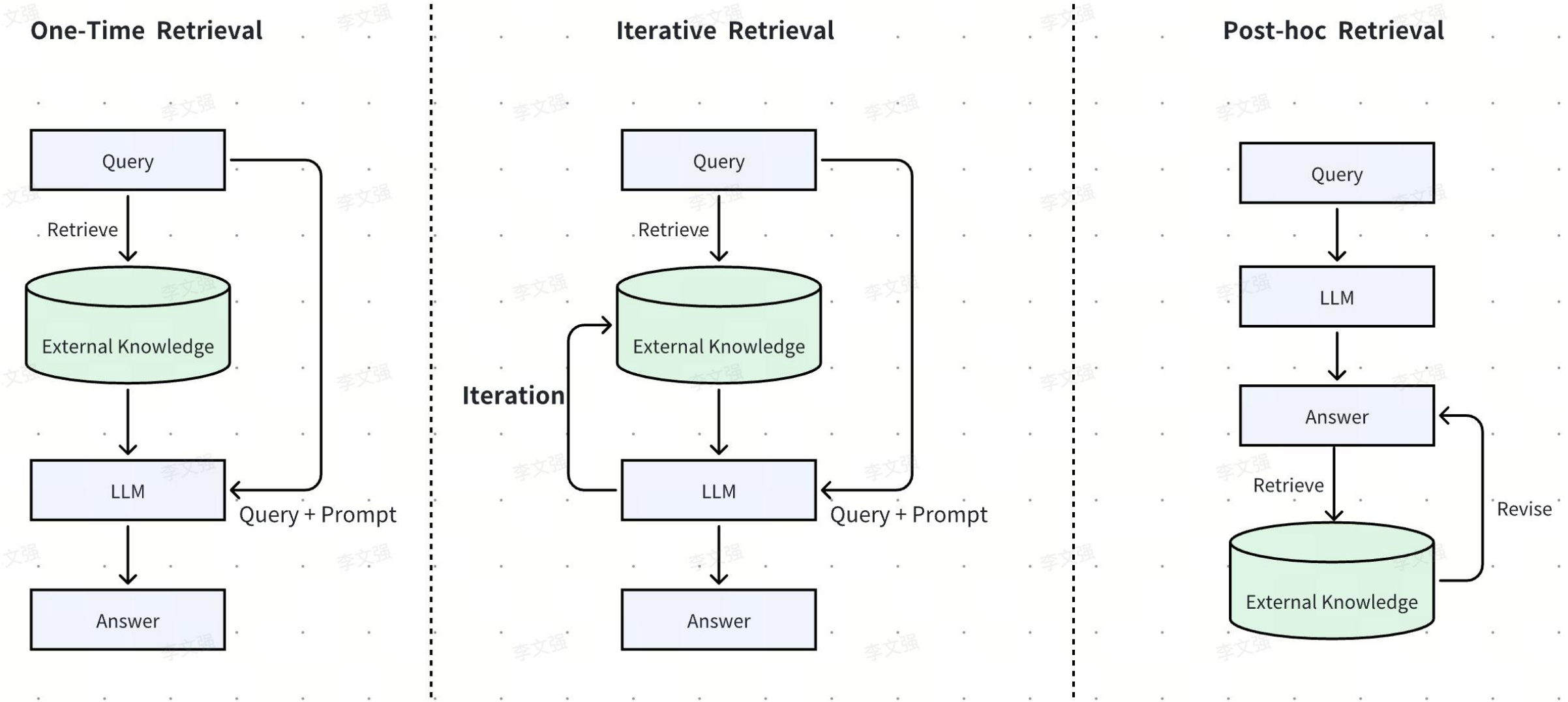
Iterative Retrieval

Retrieve external knowledge iteratively in a single task.

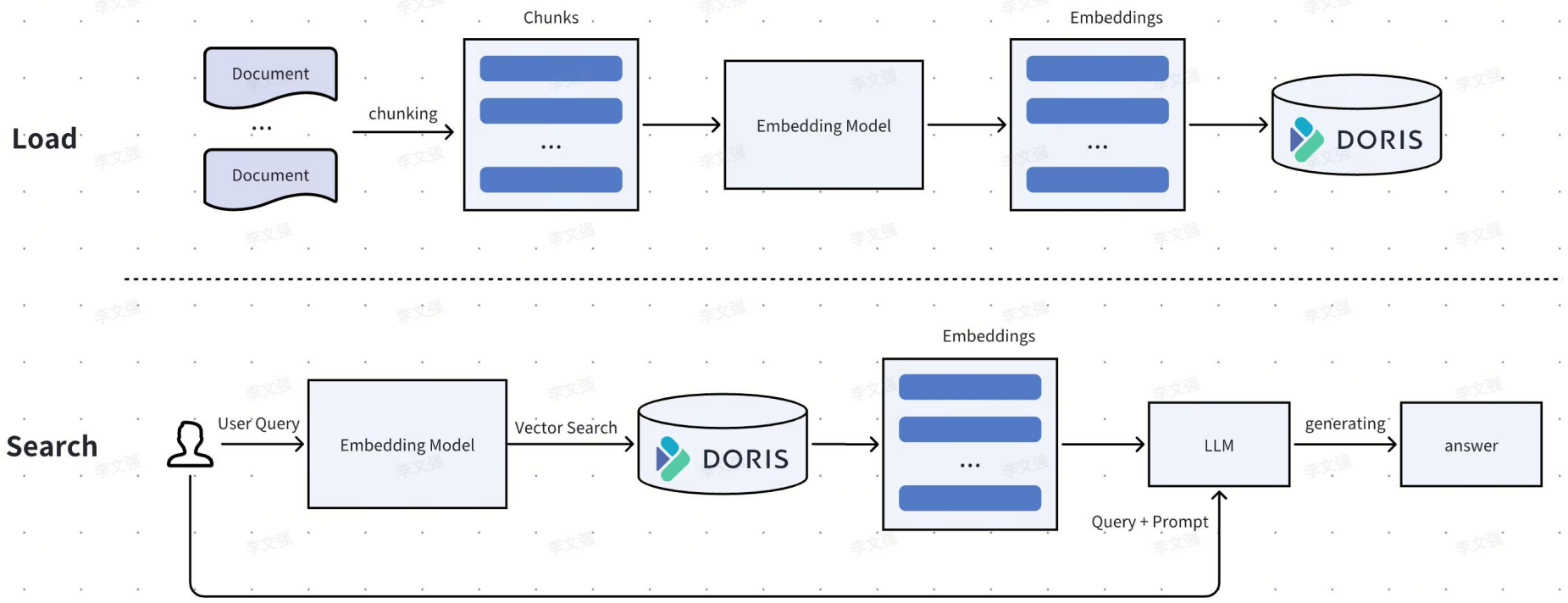
Post-hoc Retrieval

Retrieves external knowledge after generating to do revising.

Retrieval-Augmented Generation



Apache Doris 4.0 Retrieval-Augmented Generation



Database:
Doris

Data:
embeddings

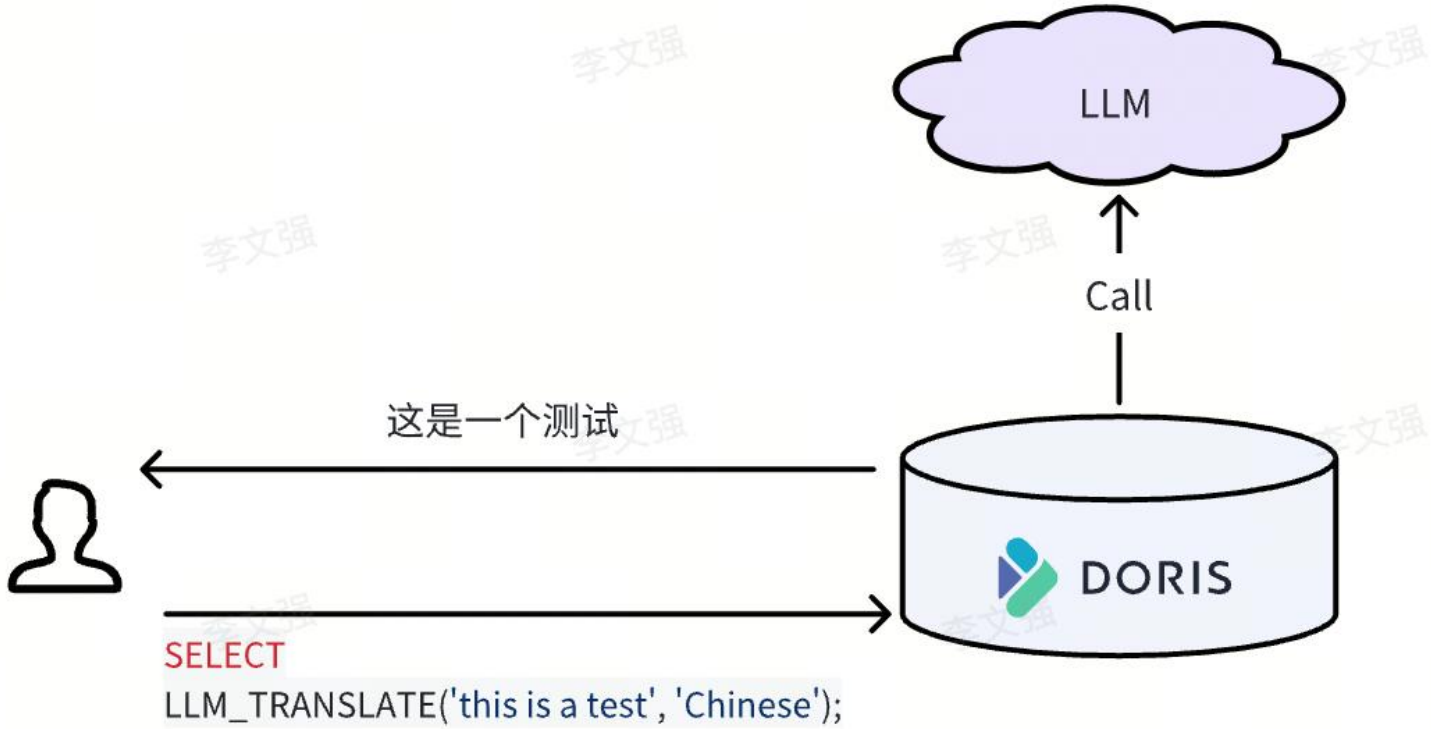
Interface:
MCP

Query type:
Hybrid/Vector searching

AI Function

AI is the **first citizen** in Apache Doris 4.0

Users could combine the **strength of AI and data analysis**



Why Apache Doris is the Answer?

High-Concurrency Test

Set-up

VeloDB: 128 cores

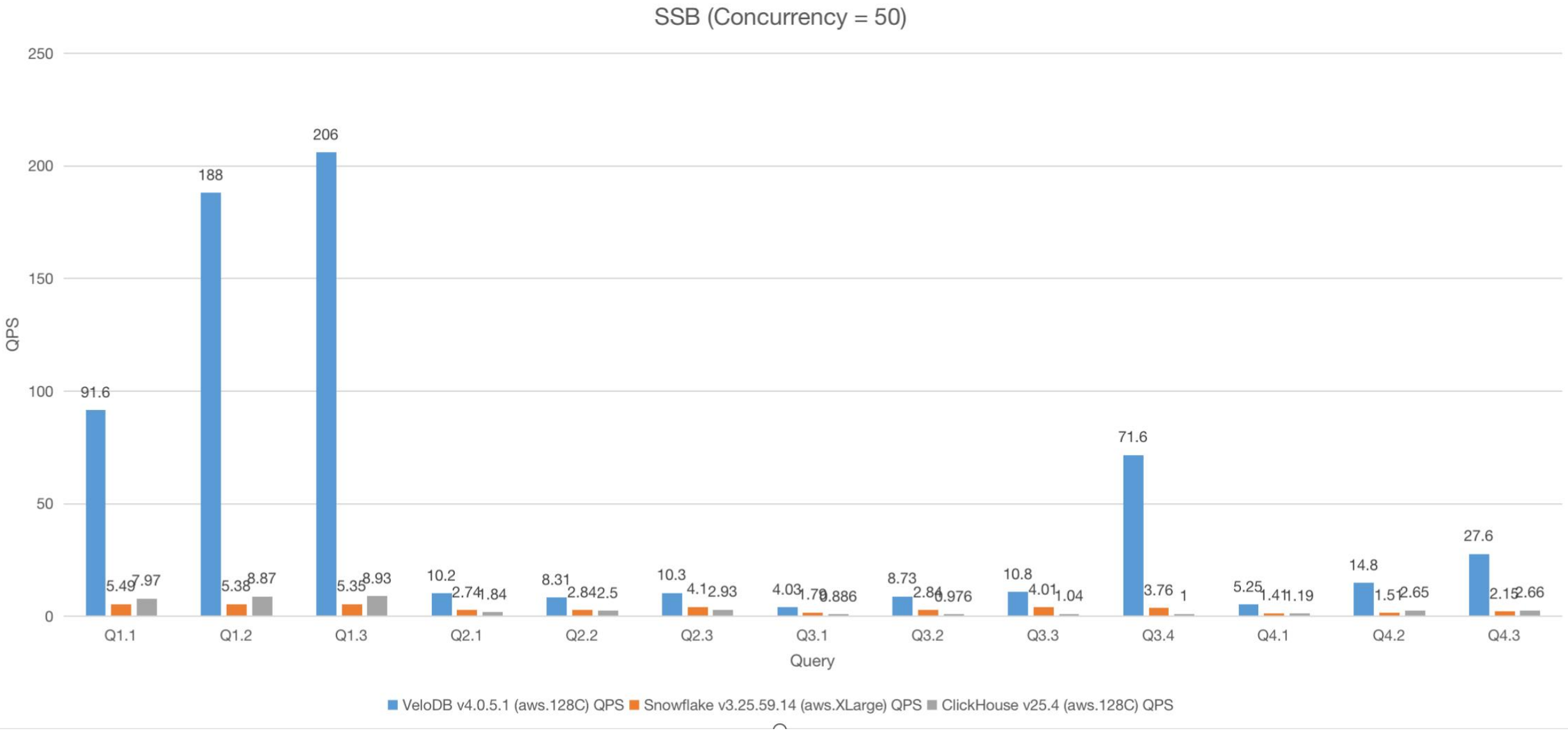
ClickHouse Cloud: 128 cores

Snowflake: X-Large Cluster

Concurrency: 50

Tool & Connection: **JMeter** testing using **JDBC** connection

High-Concurrency Test



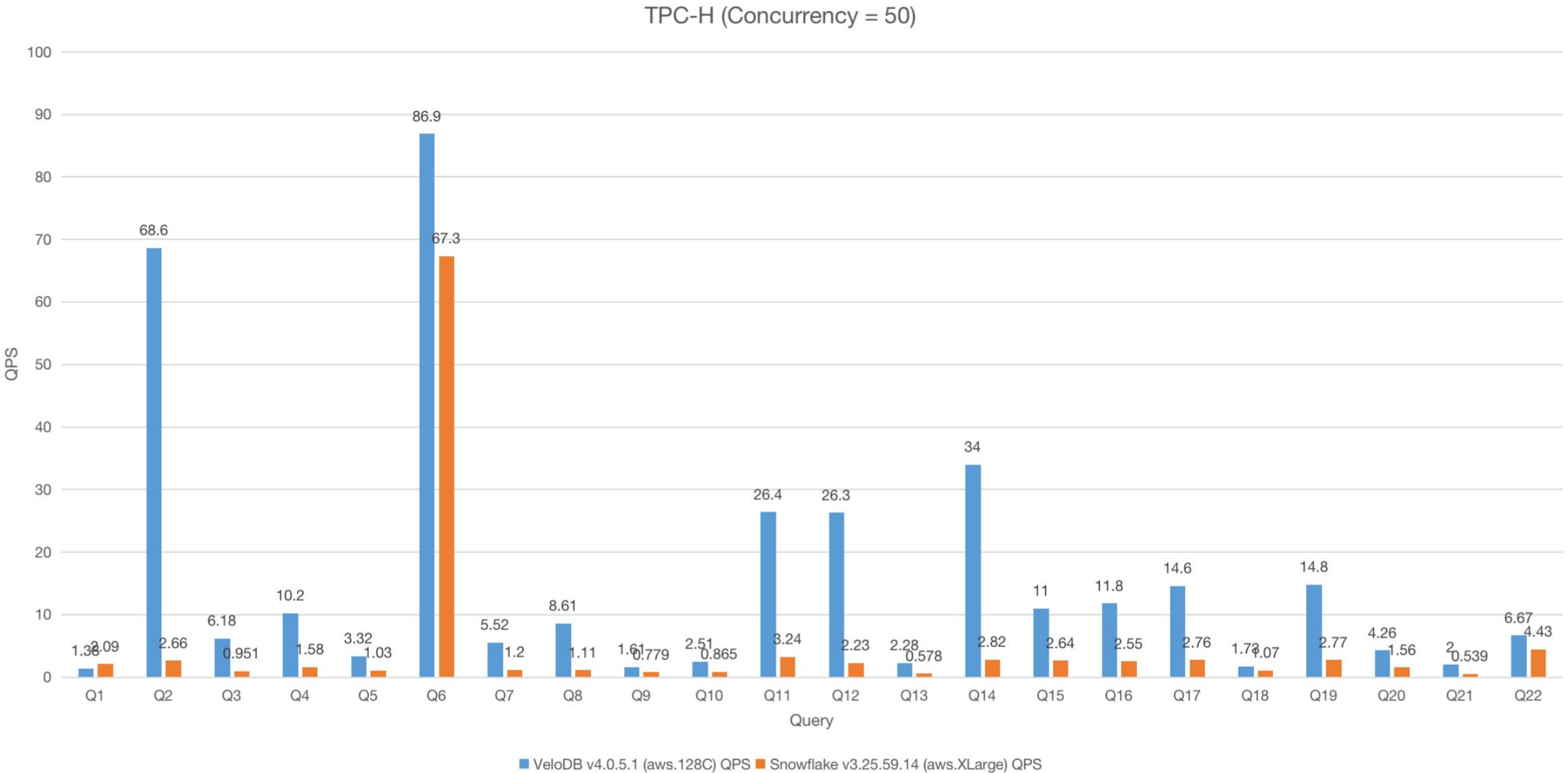
High-Concurrency Test

Avg QPS
VeloDB: 50.1
Snowflake: 3.28 (**0.07x** lower)
Clickhouse Cloud: 3.16 (**0.06x** lower)

Highlights
Q1.3
VeloDB: 206
Snowflake: 5.35 (**0.025x** lower)
Clickhouse Cloud: 8.93 (**0.043x** lower)

Q3.4
VeloDB: 71.6
Snowflake: 3.76 (**0.052x** lower)
Clickhouse Cloud: 1 (**0.013x** lower)

High-Concurrency Test



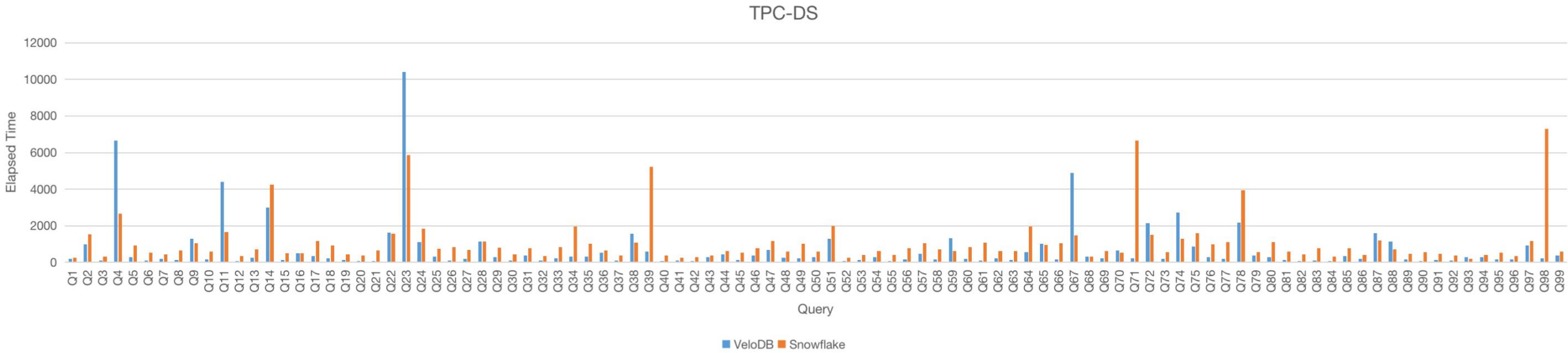
High-Concurrency Test

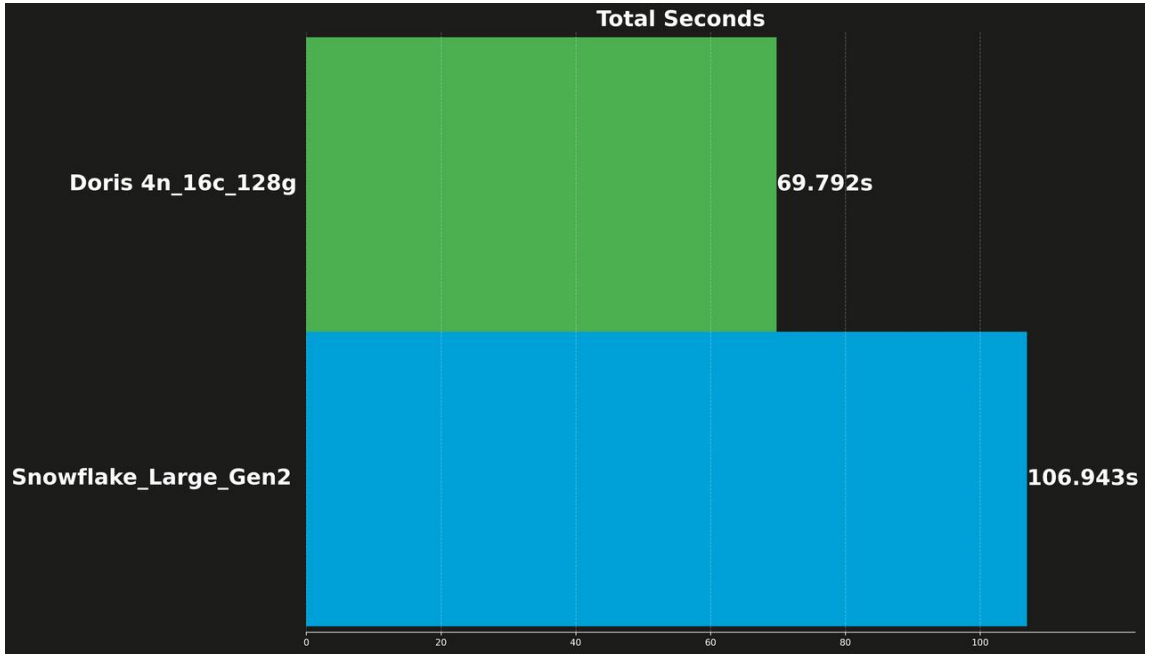
Avg QPS
VeloDB: 15.7
Snowflake: 4.31 (**0.27x** lower)
Clickhouse Cloud: **99% failure**

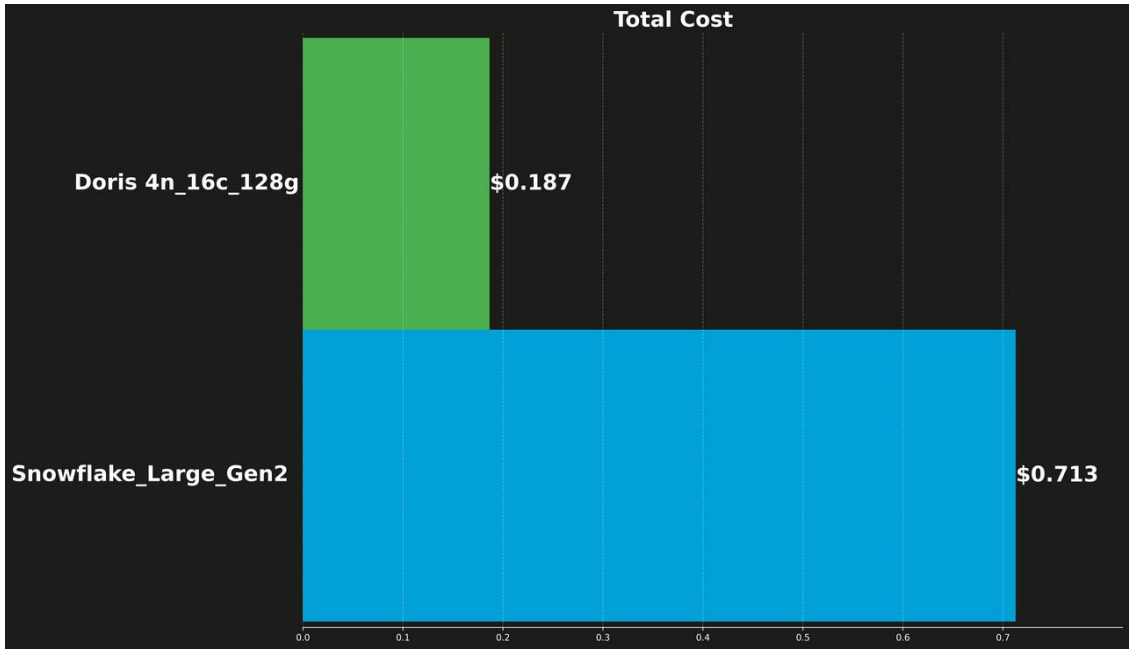
Highlights
Q2
VeloDB: 68.6
Snowflake: 2.66 (**0.038x** lower)

Q4
VeloDB: 10.2
Snowflake: 1.58 (**0.155x** lower)

Complex Queries Test







Complex Queries Test

Total Time
VeloDB: 69.792s
Snowflake: 106.943s (**1.53x** lower)
Clickhouse Cloud: **21/99 queries failure**

Costs
VeloDB: $0.187
Snowflake: $0.713 (**3.81x** higher)

Deep Diving into Apache Doris 4.0

Data Pruning

How to process data efficiently?

Not process data at all.

Static Pruning

- Predicate Filtering

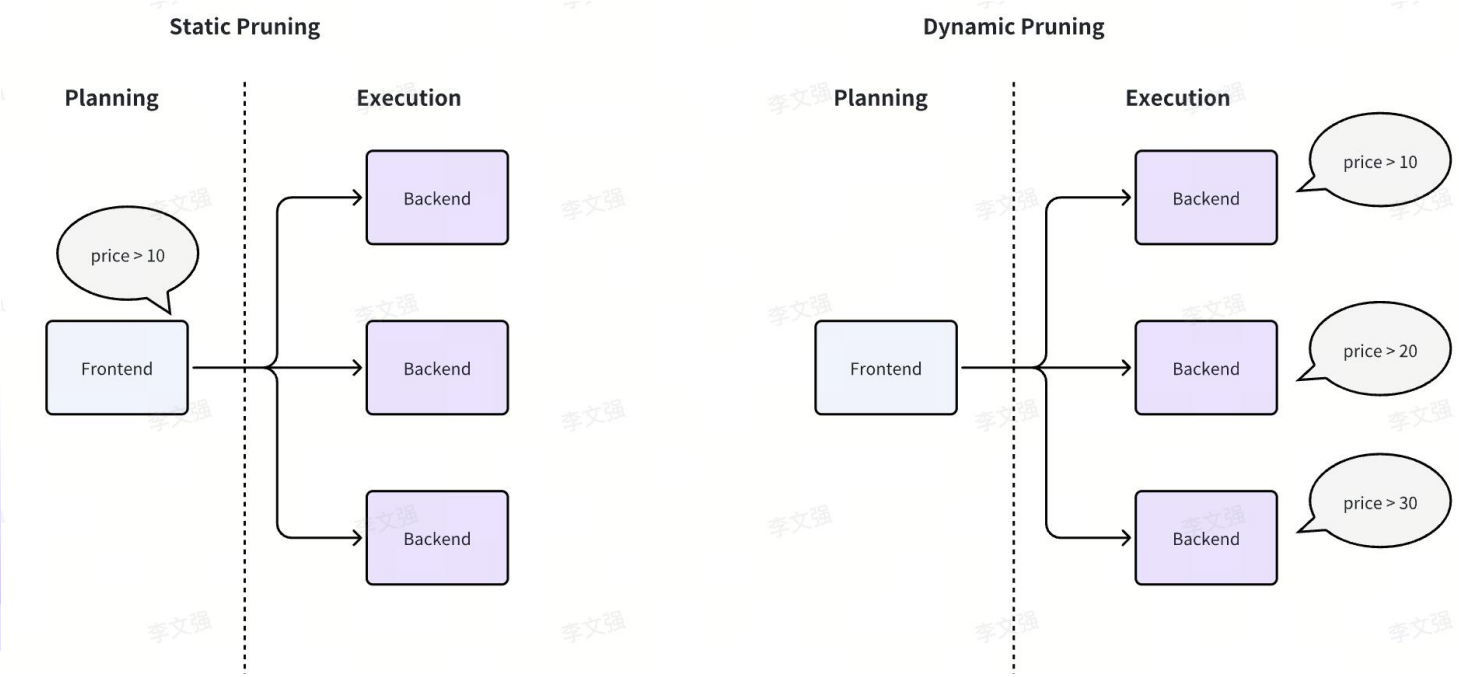
Dynamic Pruning

- LIMIT Pruning
- TopK Pruning
- Join Pruning

Data Pruning

How to process data efficiently?

Not process data at all.



1

Partition columns' predicates

Pruning based on the metadata

2

Key columns' predicates

Pruning based on upper and lower bounds (key columns are sorted in segments)

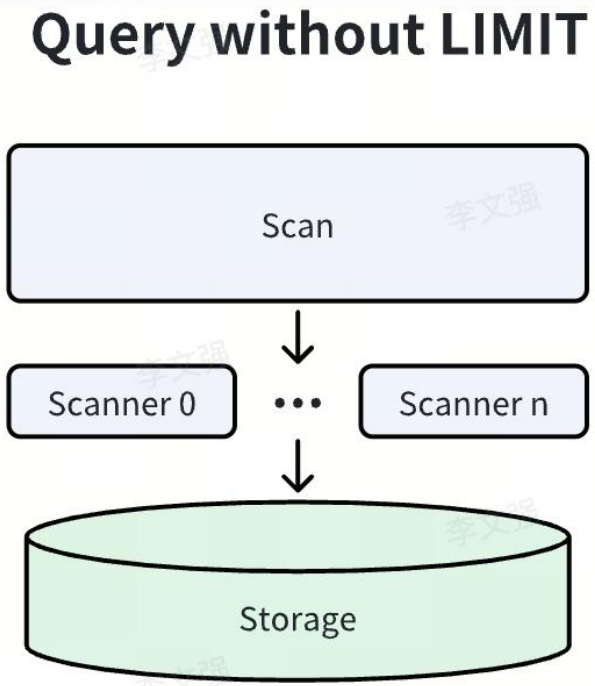
3

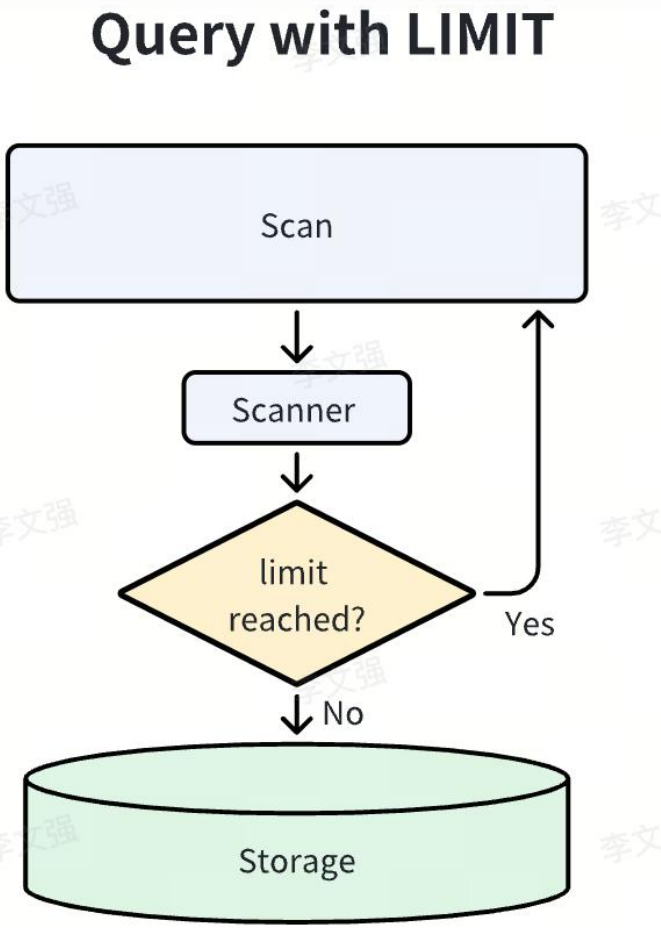
Value columns' predicate

Comparing predicate conditions and metadata

Predicate Filtering

LIMIT Pruning





Stop the query once the limitation reached

TopK Pruning

SELECT key, value FROM tbl ORDER BY key LIMIT 10;

Full Sorting + LIMIT 10

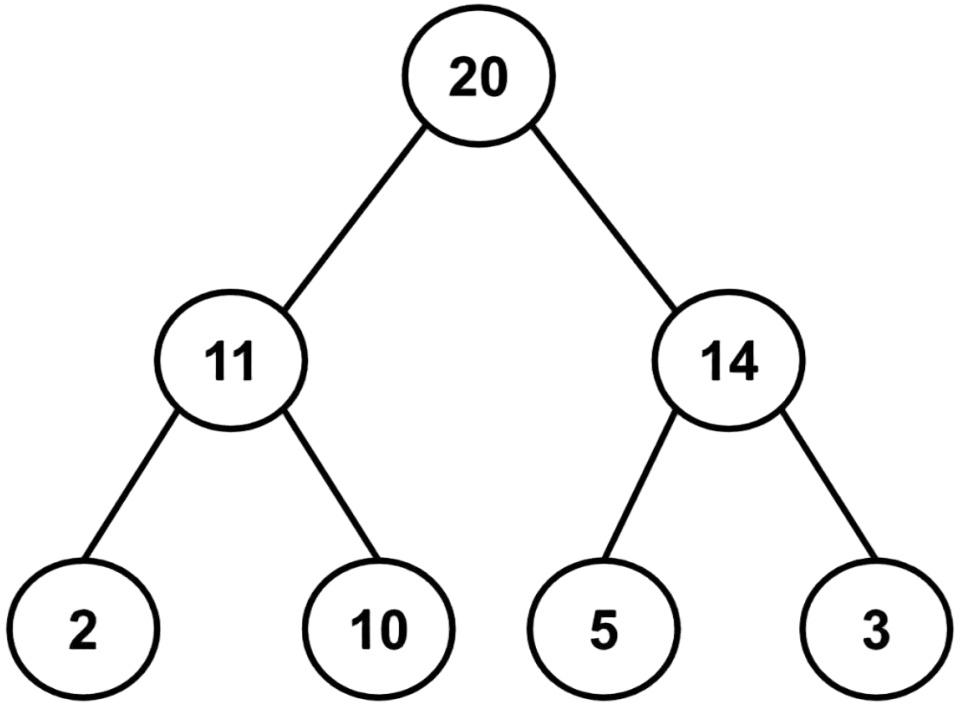
Heap Sorting
(Heap Size = 10)

TopK Pruning

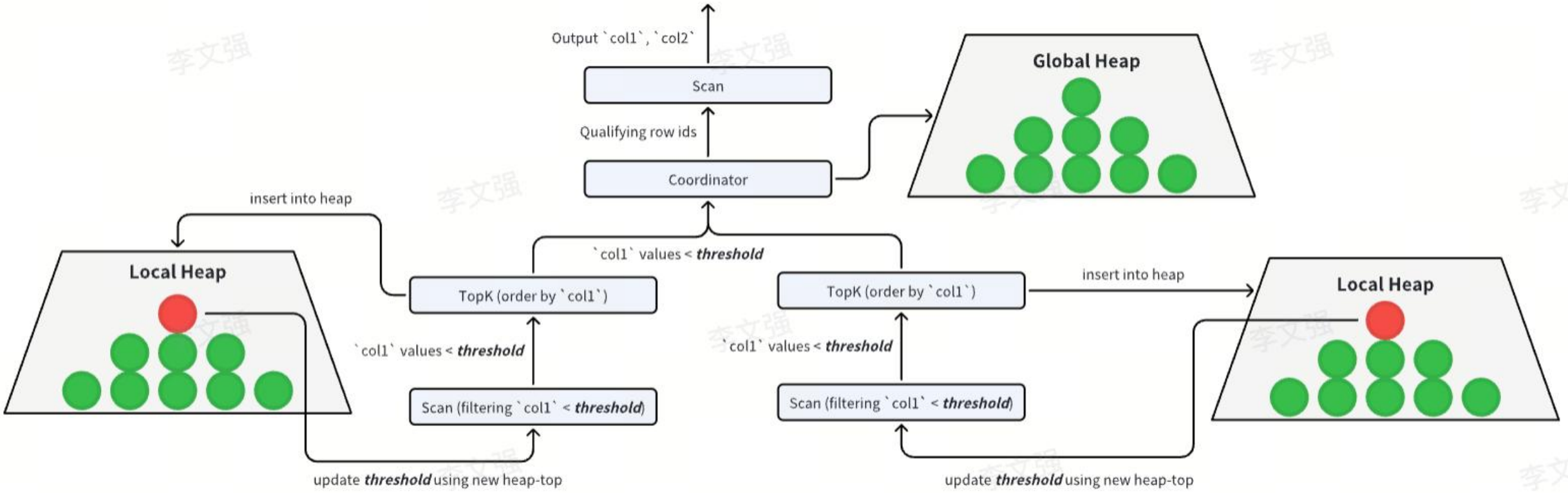
Standard Heap Sort Method
Use a min/max heap

Theoretically Optimal Solution
1 The minimum subset of data required to obtain the correct results

In Apache Doris
2 Read the first K rows of each segment and then perform a merge sort



TopK Pruning in Apache Doris



SELECT `col1`,`col2` FROM `tbl` ORDER BY `col1` LIMIT K

Join Pruning

Insights

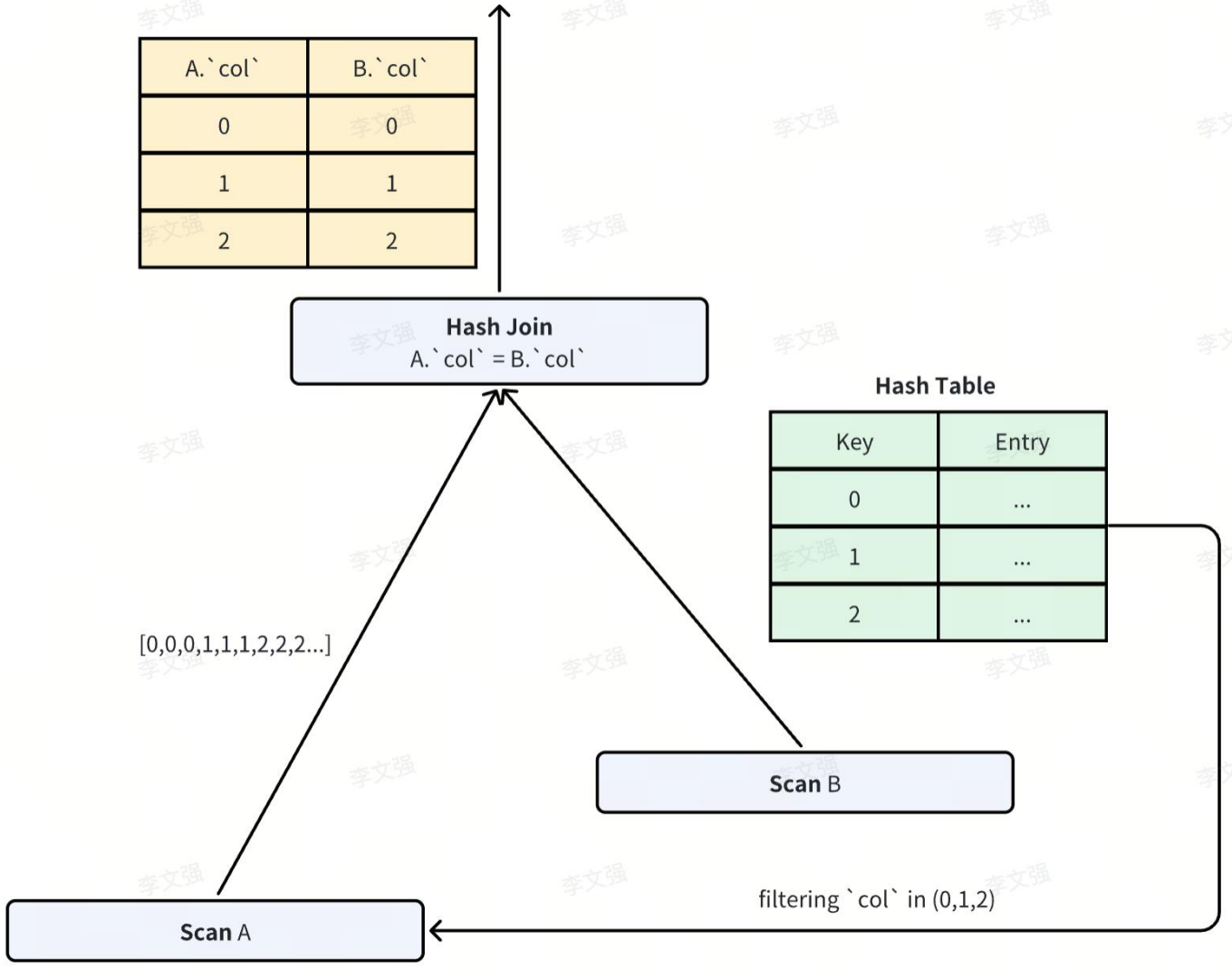
Brute-force join: O(M*N)

Hash Join: O(M+N)

Less data , lower overhead

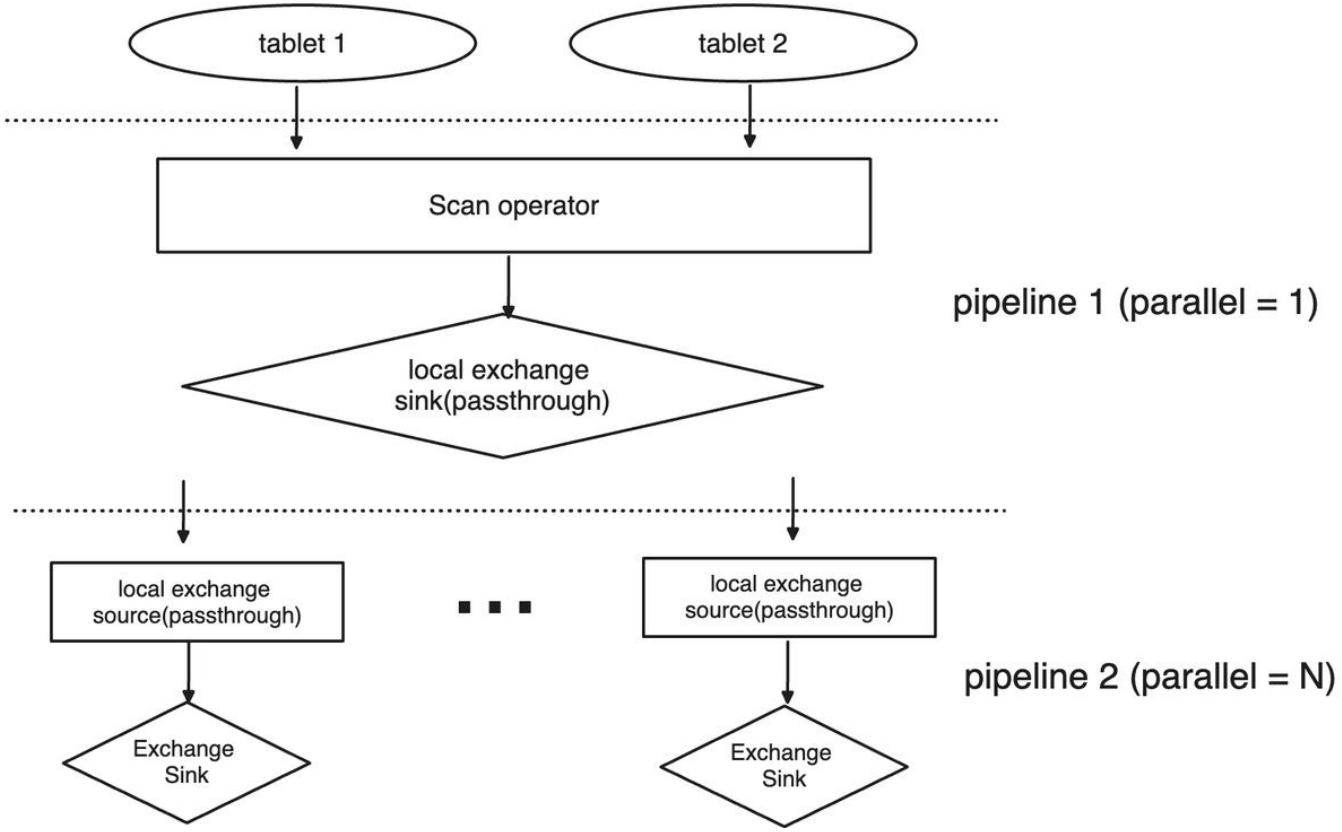
Goal: Reduce `M` from probe size

Solution: Use Join Pruning

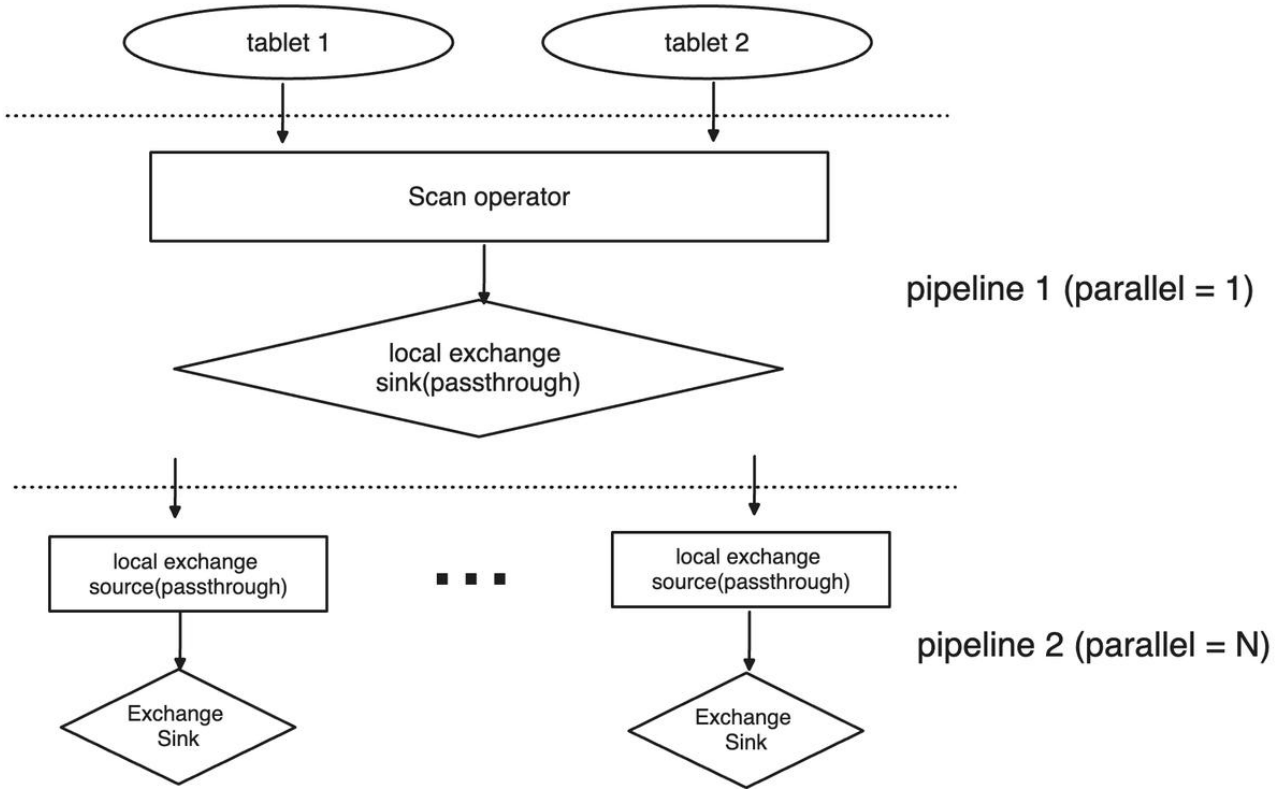


Goals:

1. Yield CPU once a blocking operator met
2. Reduce the switch overhead
3. Re-distributed data between pipelines



Pipeline Engine



Advantages:

01 Fully utilize CPU resources and distribute data evenly

02 Eliminate additional initialization overhead

03 The blocking conditions of all pipeline tasks are encapsulated using a dependency

04 Users can get a more intuitive profile

Welcome to Doris Community

Subscribe

Subscribe to our mailing list and join our discussion

dev@doris.apache.org

Get technical support

- **Slack:** apachedoriscommunity.slack.com
- **Wechat Group**: Scan the QR code on the right.

Doris

Doris AI

Thank you!