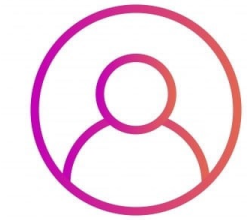
41st IEEE International Conference
on Data Engineering
HONG KONG SAR, CHINA | MAY 19 – 23, 2025

DatA Agent: A Holistic Architecture for Orchestrating Data+AI Ecosystems

Guoliang Li

Department of Computer Science, Tsinghua University

Data+AI Lifecycle



Data Inference System

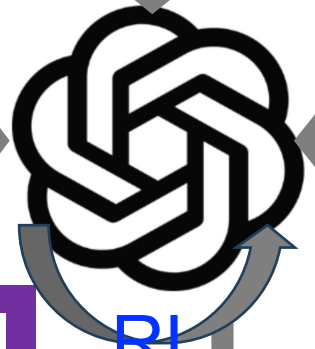
Inference Data Preparation

Data Lake

Model Inference

Training Data Preparation

Model Training



Fine-Tuning

Fine-tuning Data Preparation

Data Flywheel

RL

Rely on labor-intensive coding, hard for Adaptability ☹

Manual Intervention → Autonomy

Data + AI Motivation

How to Bridge the Gaps?

Data

Data/Query Understanding Gaps

Lack of Semantic Understanding

NLP & Semantics →

Semantic Data Organization
Semantic Query Parsing

Task Reasoning Gaps

Absence of Reasoning

Planning →

Pipeline Orchestration

Adaptability Gaps

Limited Adaptability

Reasoning
Generation →

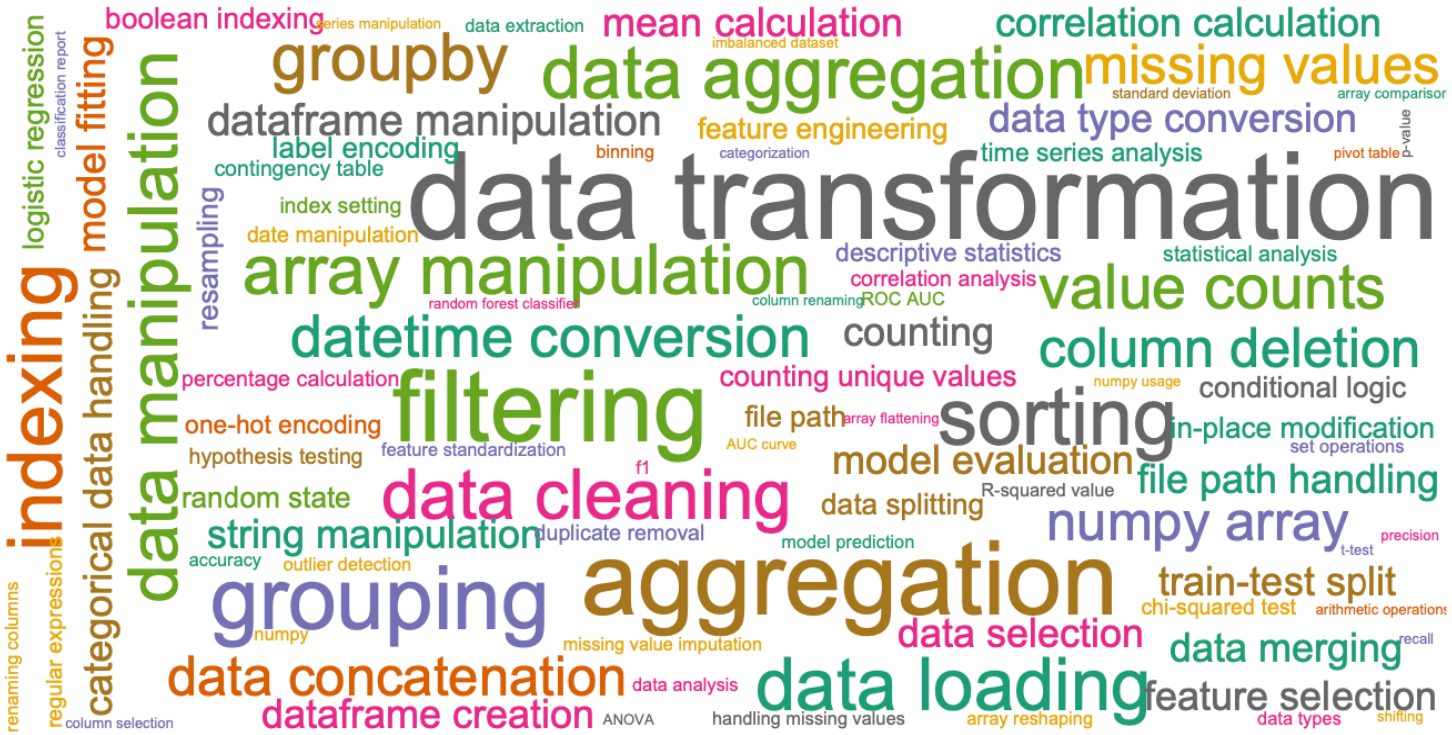
LLM/Code/Tool Hybrid Execution

An autonomous system is crucial for Data+AI applications.

Data Agent

Data Agent: Autonomously performs data-related tasks with capabilities for knowledge comprehension, automatic planning, and self-reflection.

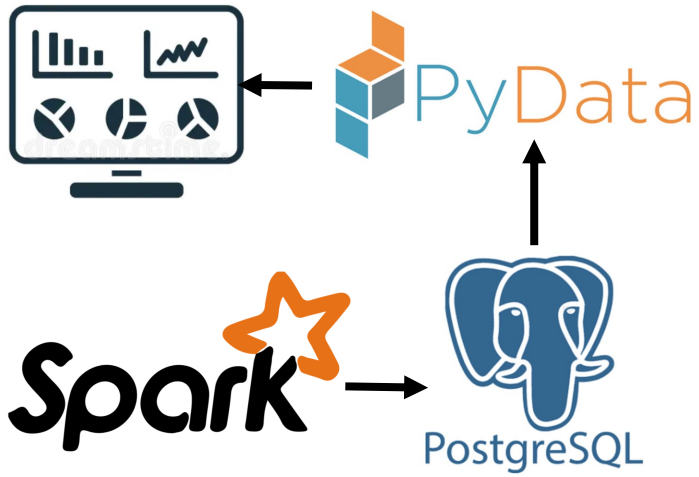
- Data Analytics Agent
  - Unstructured Data Agent
  - Semantic Structured Data Agent
  - Data Lake Agent
  - Multi-Modal Data Agent
- Data Science Agent
- DBA Agent
- Database Development Agent



Data Agent: Framework

NL Queries

Answers



Pipeline Orchestration Agents

Semantic Catalog

Data Explora-tion Agents

VDB

Engine Schedule Agents

Engine Catalog

Data Fabric Tools

Data Access Tools

Data Processing Tools

Data Viz Tools

a holistic view of diverse sources to unify data management

Data Lake

Data Agent: Architecture

Applications

Chat2BI | NL2SQL | TableQA | Report Generation | Semantic Analytics | Multi-Modal Analytics

Orchestration Plane

NL2Pipeline | Pipeline Orchestration | Pipeline Optimization | Pipeline Execution | Pipeline Reflection

Engine Plane

Agents Modeling | Agents Scheduling | Agents Interaction | Agent-Tool Scheduling | Agent-Tool Interaction

Data Plane

Semantic Meta Data | Semantic Organization | Semantic Catalog | Semantic Exploration | Semantic Indexing

Data Agent: Challenges

Challenge ①: How to understand queries, data, agents, tools?

Challenge ②: How to orchestrate an effective and efficient pipeline?

Challenge ③: How to schedule and interact agents and tools?

Challenge ④: How to optimize/execute a pipeline?

Challenge ⑤: How to continuously enhance the pipeline quality?

Data Agent: Challenges

- Data/Query Understanding
- Environment Understanding
- Optimization Goal

Perception

- Task Decomposition
- Task Planning
- Task Optimization
- Task Execution

Reasoning & Planning

- Tool Understanding
- Tool Selection
- Tool Interaction
- Tool Standardization

Tool Invocation

- Domain Knowledge
- Context Knowledge
- Environment Knowledge
- Reflective Knowledge

Memory

- Reflection
- RL
- Reward Model

Continuous Learning

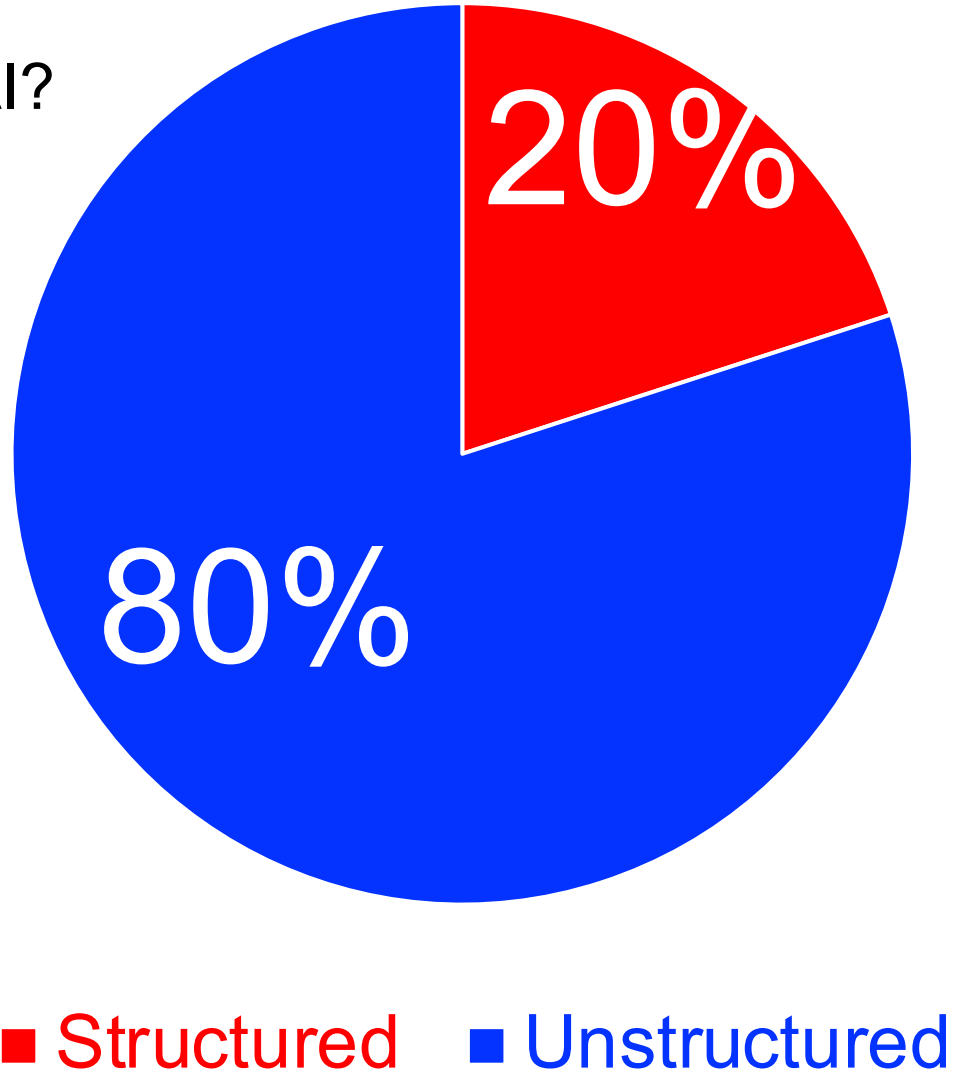
- Agent Card
- Agent-Agent: A2A
- Agent-Tool: MCP
- Agent-Network: ANP

Multiple Agents

Data Agent for Unstructured Data Analytics

Motivation of Unstructured Data Analytics

- Hard to Analyze Unstructured Data!
  - How many ICDE25 papers are related to Data+AI?
  - Which conferences have the highest number of Data+AI papers?
- Challenges
  - Query Understanding
  - Unstructured Data Filtering (Operators)
  - Unstructured Data Aggregation (Operators)
- Autonomous Unstructured Data Analytics Pipeline Orchestration & Optimization



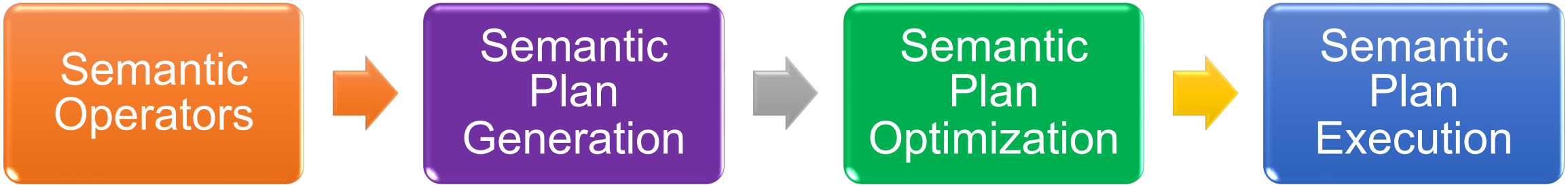
Multi-hop Query Example

List the highest-rated movies directed by Steven Spielberg before he turned 40.

Step 1. **Semantic Extract** the year of Steven Spielberg' 40 years old

Step 2. **Semantic Filter** the movies before this year

Step 3. **Semantic Rank** the movies by the ratings



Semantic Data Analytics Example

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

Step 1. **Semantic Extract** movies directed by Steven Spielberg

Step 2. **Semantic Filter** the reports for the extracted movies

Step 3. **Semantic Group** the reports by filtered movies

Step 4. **Semantic Classify** the reports as Negative or Positive

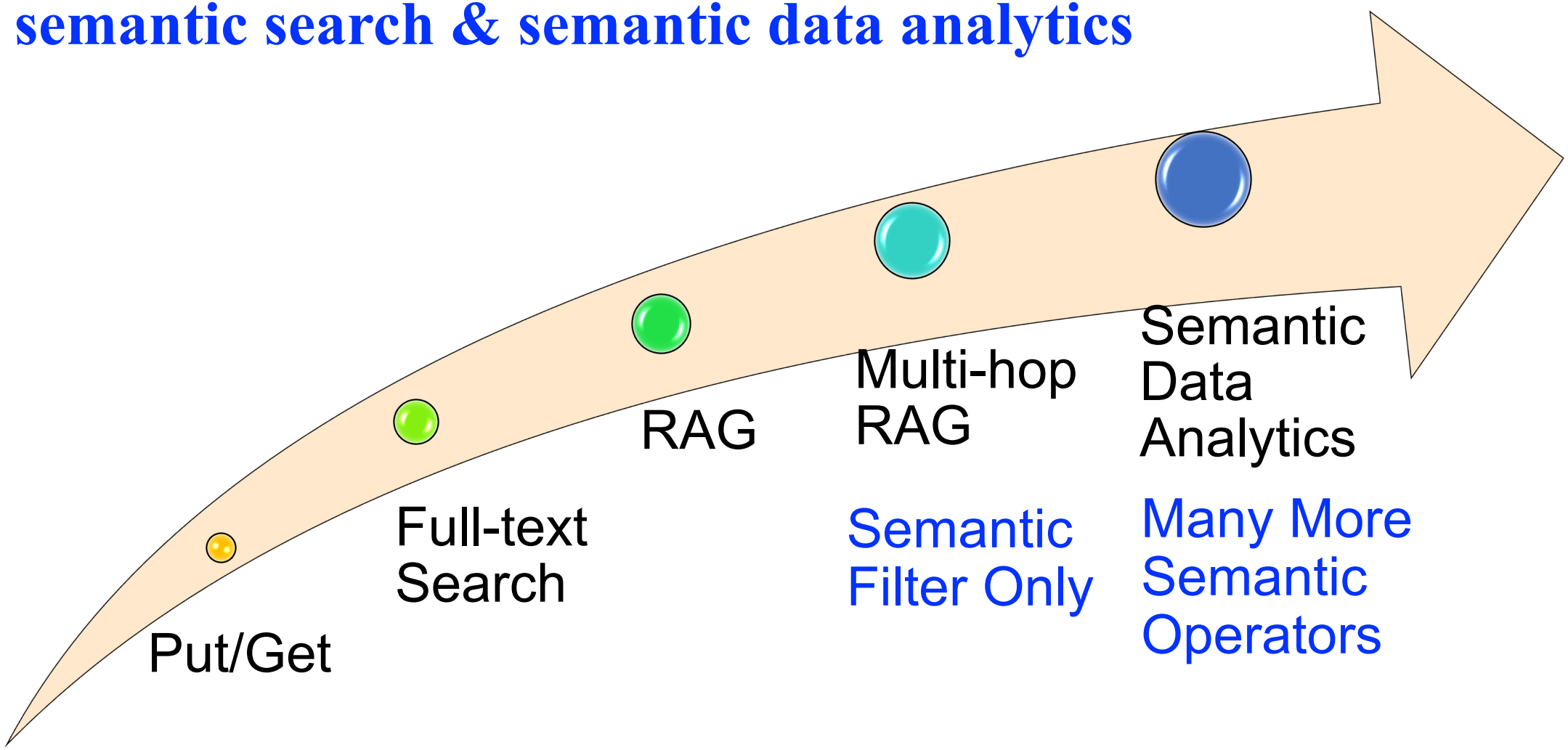
Step 5. **Count** the Positive reports and Negative reports for each group

Step 6. **Compare** #Positive reports and #Negative reports

Step 7. **Count** the number of groups with more #Positive reports

Unstructured Data Analytics Agent

Semantic Unstructured Data Analytics Agent for multi-hop semantic search & semantic data analytics



Jiayi Wang, Guoliang Li. AOP: Automated and Interactive LLM Pipeline Orchestration for Answering Complex Queries. CIDR 2025

Unstructured Data Analytics Agent: NL2Pipeline

NL Query

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

① Semantic Extract
② Semantic Filter
③ Semantic Group
④ Semantic Classify
⑤ Count
⑤ Count
⑥ Compare
⑦ Count

Semantic Extract | Semantic Filter | Semantic Group | Semantic Classify | Count | Compare

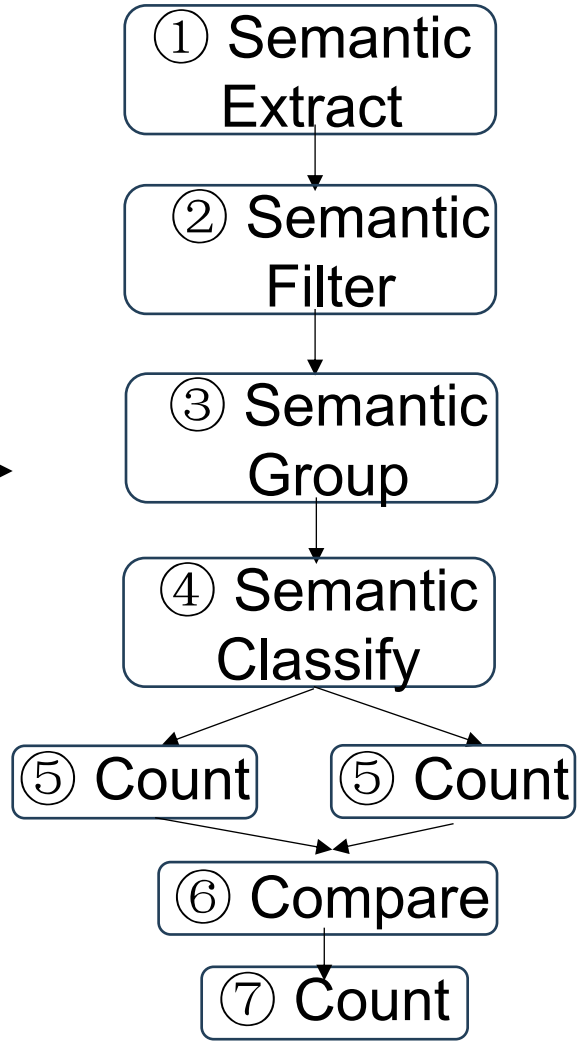
Unstructured Data Analytics Agent: NL2Pipeline

NL Query

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

LLM Capabilities

- NL Understanding
- Semantic Matching
- Reasoning
- Generation



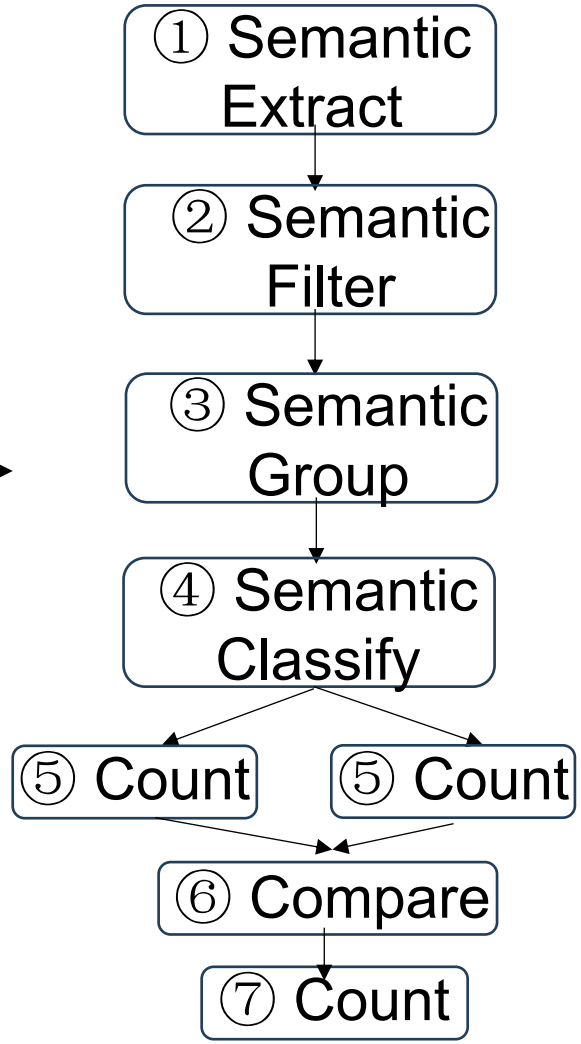
Unstructured Data Analytics Agent: NL2Pipeline

NL Query

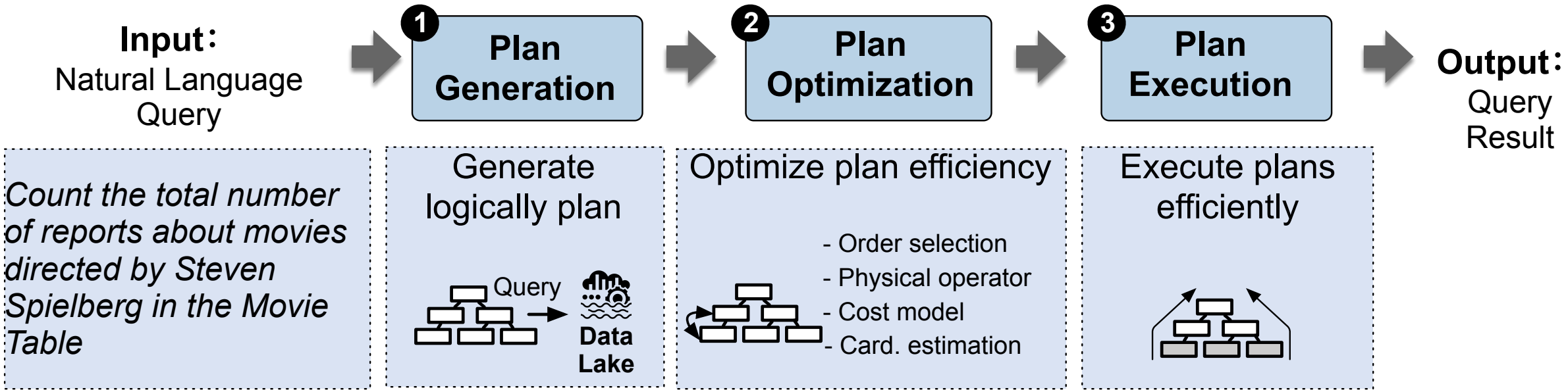
Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

Challenges

- Semantic Operators
- Pipeline Orchestration
- Pipeline Optimization
- Pipeline Execution

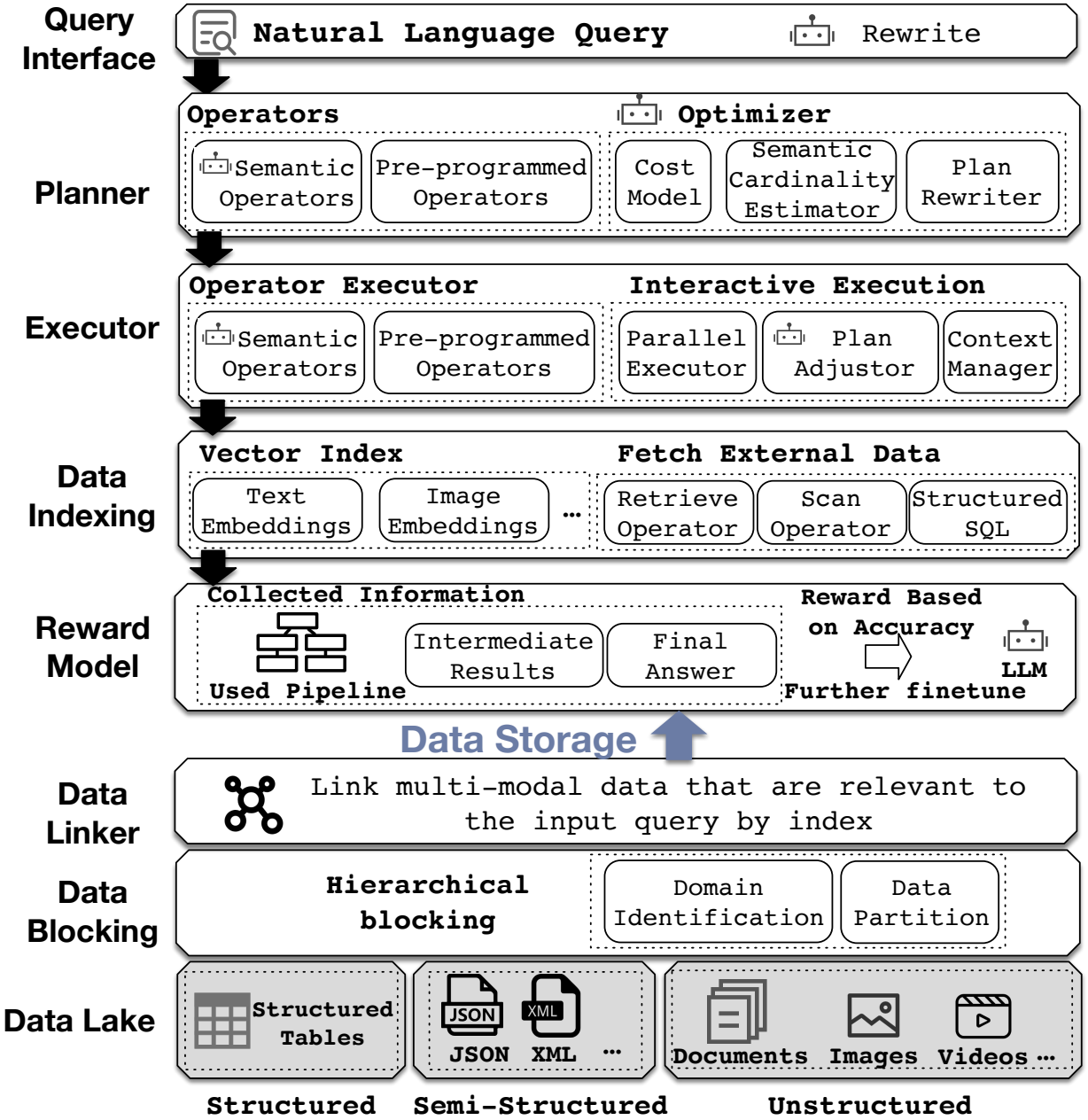


Unstructured Data Analytics Agent : Overview



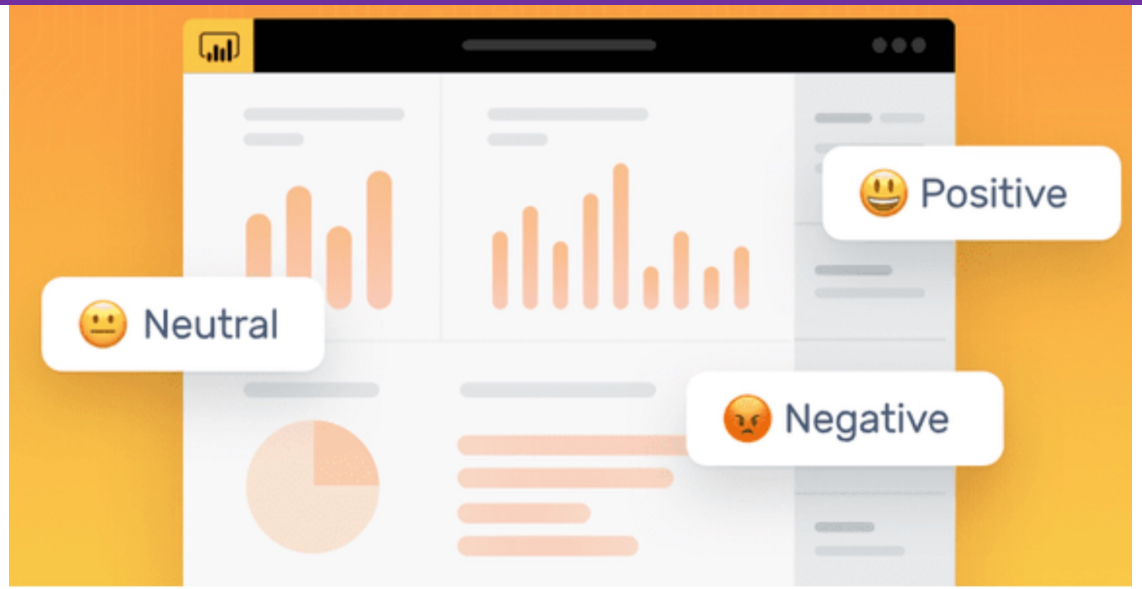
Unstructured Data Analytics Agent : Architecture

- Semantic Query: NL
- Semantic Operators:
  - LLM Semantic Operators
  - Programmed Operators
- Semantic Planning:
  - Pipeline orchestration
- Semantic Optimization & Execution
  - Multi-goal optimization
  - Interactive execution
- Self-reflection
  - COT, Reward model
- Semantic Indexing
  - Vector indexes



Unstructured Data Analytics Agent: Semantic Operators

- Extension from DB Operators
  - Semantic filter
  - Semantic projection
  - Semantic join
  - Semantic group-by
  - Semantic order-by
- Unstructured Operators
  - Semantic extract
  - Semantic classify
  - Semantic segmentation



Semantic Operators: Two Examples

✓ **Semantic Filter**

- **Logical Representations:**
  - [Entity] satisfy [Condition]
  - [Entity] that [Condition]
- **Physical implementations:**
  - ① Keyword filtering
  - ② Embedding-based filtering
  - ③ LLM filtering

✓ **Semantic GroupBy**

- **Logical Representations:**
  - Group [Entity] by [Attribute]
  - For each [Entity], aggregate
  - The number of [Entity]
- **Physical implementations:**
  - ① Hash-based Groupby
  - ② Embedding-based Groupby
  - ③ LLM-based Groupby

Logical: Fixed Reserved Words → Semantic Words (for NL queries)

Physical: Fixed implementation → Semantic implementation

Semantic Operators: Definition

- Relational operators are insufficient for unstructured data analytics
  - Lack semantic processing capabilities
- To address this, we manually identify a set of operators to support more comprehensive analytics, each defined with:
  - Input
  - Output
  - Predefined execution processes (physical operators)
  - A set of logical representations
- **A logical representation** is **a structured template that abstracts the semantic essence of NL expressions** into placeholders like Entity and Condition, capturing distinct semantic roles.
  - Filter: [Entity] satisfy [Condition]; [Entity] that [Condition]; [Entity] by [Condition]

Semantic Operators: Logical/Physical Implementation

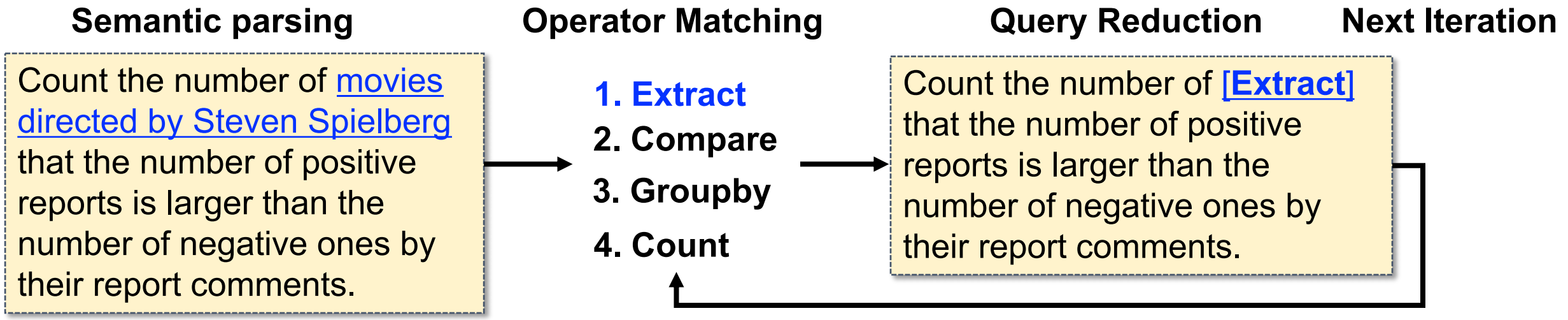
- **Logical operators: Filter, Extract, Compare, Group By, Classify, Order By …**
- **Physical operators: Pre-programmed, LLM-based, LLM-based UDF**

THE LOGICAL OPERATORS, THEIR INPUTS, OUTPUTS, CORRESPONDING PHYSICAL OPERATORS, AND EXAMPLE LOGICAL REPRESENTATIONS.

Operator	Input	Output	Pre-programmed Implementation	LLM-based Implementation	Example Logical Representation
`Scan`	List	List	Linear Scan, Index Scan	-	documents satisfy [Condition]
`Filter`	List	List	Exact condition filtering	Semantic filtering	[Entity] that [Condition]
`Compare`	A, B, Condition	A/B	Standard comparison, *e.g.*, $>$, $<$	Semantic comparison	larger in [Entity] and [Entity]
`GroupBy`	List	List of List	Grouping by exact attributes	Semantic grouping	aggregate [Entity] by [Attribute]
`Count`	List	Number	Standard aggregation (Count)	Semantic count	number of documents [Condition]
`Sum`	List	Number	Standard aggregation (Sum)	Semantic sum	the total sum of [Entity]
`Max`	List	Number	Standard aggregation (Max)	Semantic max	the maximum of [Entity]
`Min`	List	Number	Standard aggregation (Min)	Semantic min	the minimum of [Entity]
`Average`	List	Number	Standard aggregation (Average)	Semantic average	the mean of [Entity]
`Median`	List	Number	Standard aggregation (Median)	Semantic median	the median of [Entity]
`Percentile`	List	Number	Standard aggregation (Percentile)	Semantic percentile	the k-th percentile for [Entity]
`OrderBy`	List	List	Numerical/lexicographical sort	Semantic sorting	Sort [Entity] [Condition]
`Classify`	Text	Class	Rule-based/ML-based classification	Semantic classification	The type of [Entity]
`Extract`	Text	Text	Keyword/Regex extraction	Semantic extraction	get [Entity] from documents
`TopK`	List	List	Numeric ranking	Semantic ranking	the top [Number] [Entity]
`Join`	List, List	List	Join by key	Semantic join	[Entity] that also occurs in [Entity]
`Union`	Set, Set	Set	Standard set union	Semantic set union	set union of [Entity] and [Entity]
`Intersection`	Set, Set	Set	Standard set intersection	Semantic set intersection	in set [Entity] and in [Entity]
`Complementary`	Set, Set	Set	Standard set complementary	Semantic set complementary	in set [Entity] not in [Entity]
`Compute`	List	Number	Programmed Mathematical Equation	Semantic computation	sum of squares of [Entity]
`Generate`	Text	Text	-	LLM invocation	explain the result

Pipeline Planning: Automatic Orchestration

- **Overview**: progressively identifying appropriate pre-defined logical operators and reducing the query with the operators.
  - ① Semantic Parsing: extract the logical representations from the query
  - ② Operator Matching: identify the matched logical operators
  - ③ Query Reduction: reduce the logical operators to generate a plan
  - ④ Error Handling: backtrack to the previous reduction



Pipeline Orchestration

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

Pipeline Orchestration

①

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

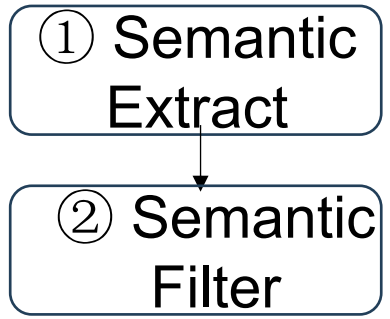
① Semantic Extract

Step 1. **Semantic Extract** movies directed by Steven Spielberg

Pipeline Orchestration

① ②

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.



Step 1. **Semantic Extract** movies directed by Steven Spielberg

Step 2. **Semantic Filter** the reports for the movies

Pipeline Orchestration

③ ①

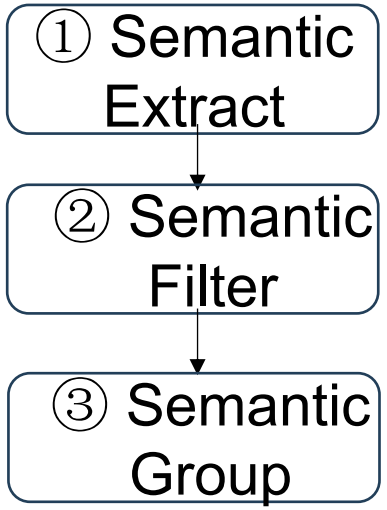
Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

②

Step 1. **Semantic Extract** movies directed by Steven Spielberg

Step 2. **Semantic Filter** the reports for the movies

Step 3. **Semantic Group** the reports by movie



Pipeline Orchestration

③ ① ④

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

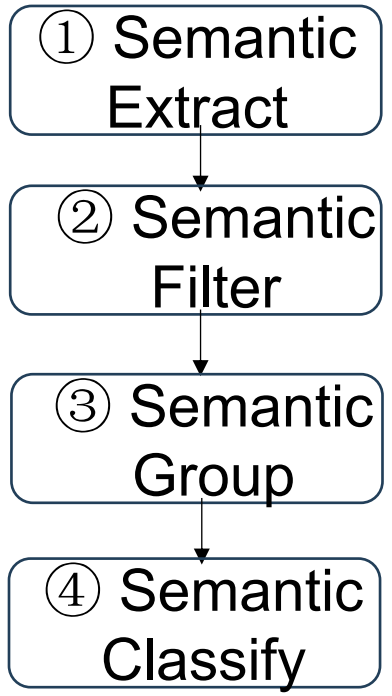
④ ②

Step 1. **Semantic Extract** movies directed by Steven Spielberg

Step 2. **Semantic Filter** the reports for the movies

Step 3. **Semantic Group** the reports by movie

Step 4. **Semantic Classify** Negative or Positive



Pipeline Orchestration

③ ① ⑤ ④

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

⑤ ④ ②

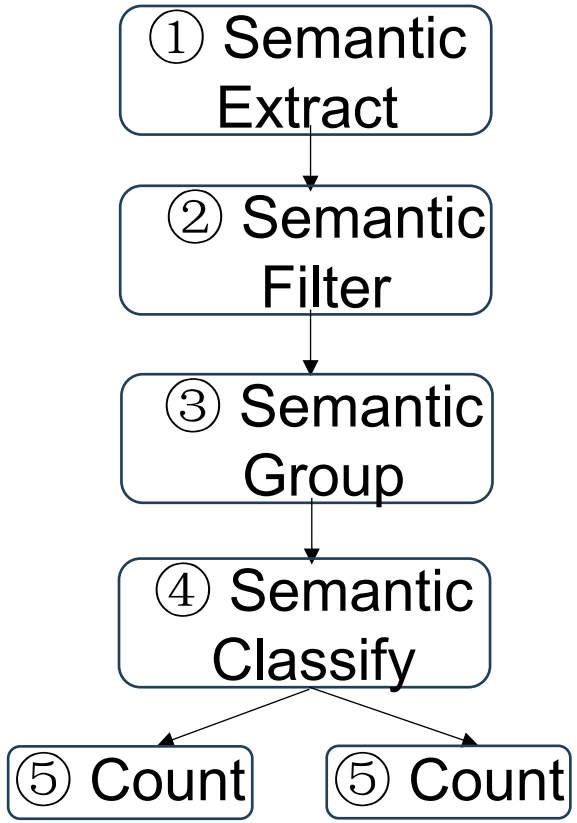
Step 1. **Semantic Extract** movies directed by Steven Spielberg

Step 2. **Semantic Filter** the reports for the movies

Step 3. **Semantic Group** the reports by movie

Step 4. **Semantic Classify** Negative or Positive

Step 5. **Count** Positive/Negative reports



Pipeline Orchestration

③ ① ⑤ ④

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

⑥ ⑤ ④ ②

Step 1. **Semantic Extract** movies directed by Steven Spielberg

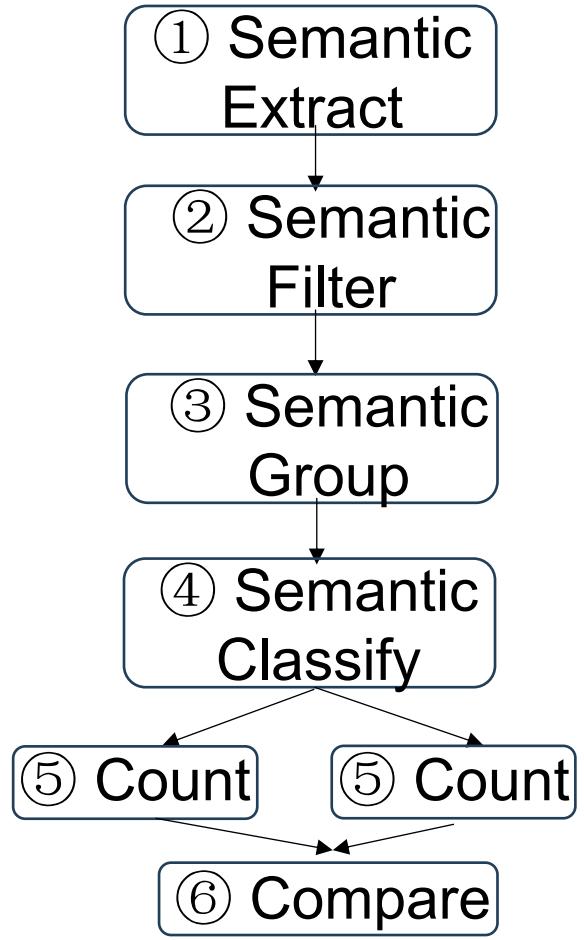
Step 2. **Semantic Filter** the reports for the movies

Step 3. **Semantic Group** the reports by movie

Step 4. **Semantic Classify** Negative or Positive

Step 5. **Count** Positive/Negative reports

Step 6. **Compare** #Positive and #Negative reports



Pipeline Orchestration

⑦ ③ ① ⑤ ④

Count the number of movies directed by Steven Spielberg that the number of positive reports is larger than the number of negative ones by their report sentiments.

⑥ ⑤ ④ ②

Step 1. **Semantic Extract** movies directed by Steven Spielberg

Step 2. **Semantic Filter** the reports for the movies

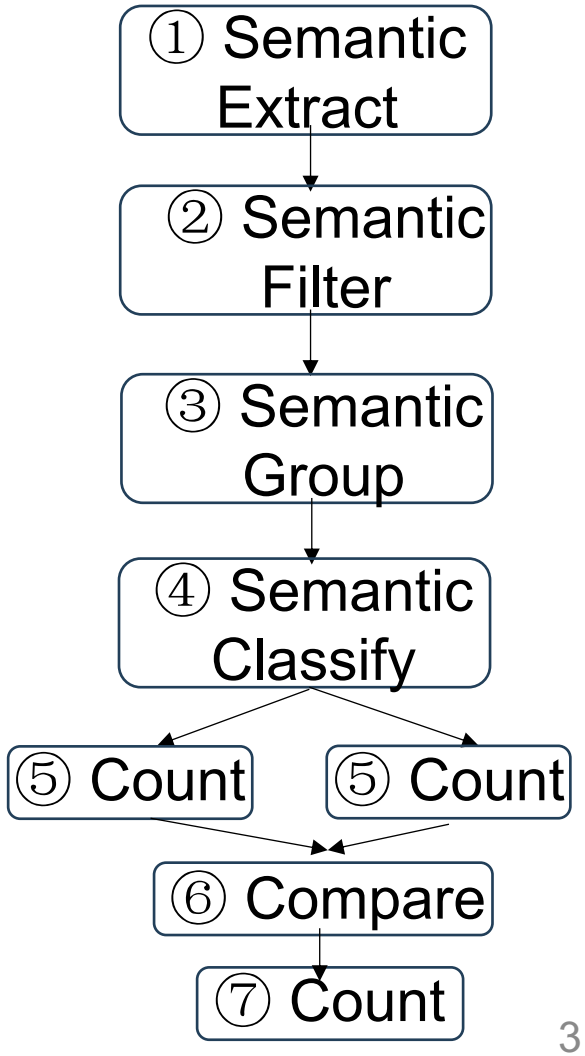
Step 3. **Semantic Group** the reports by movie

Step 4. **Semantic Classify** Negative or Positive

Step 5. **Count** Positive/Negative reports

Step 6. **Compare** #Positive and #Negative reports

Step 7. **Count** the groups with more Positive reports

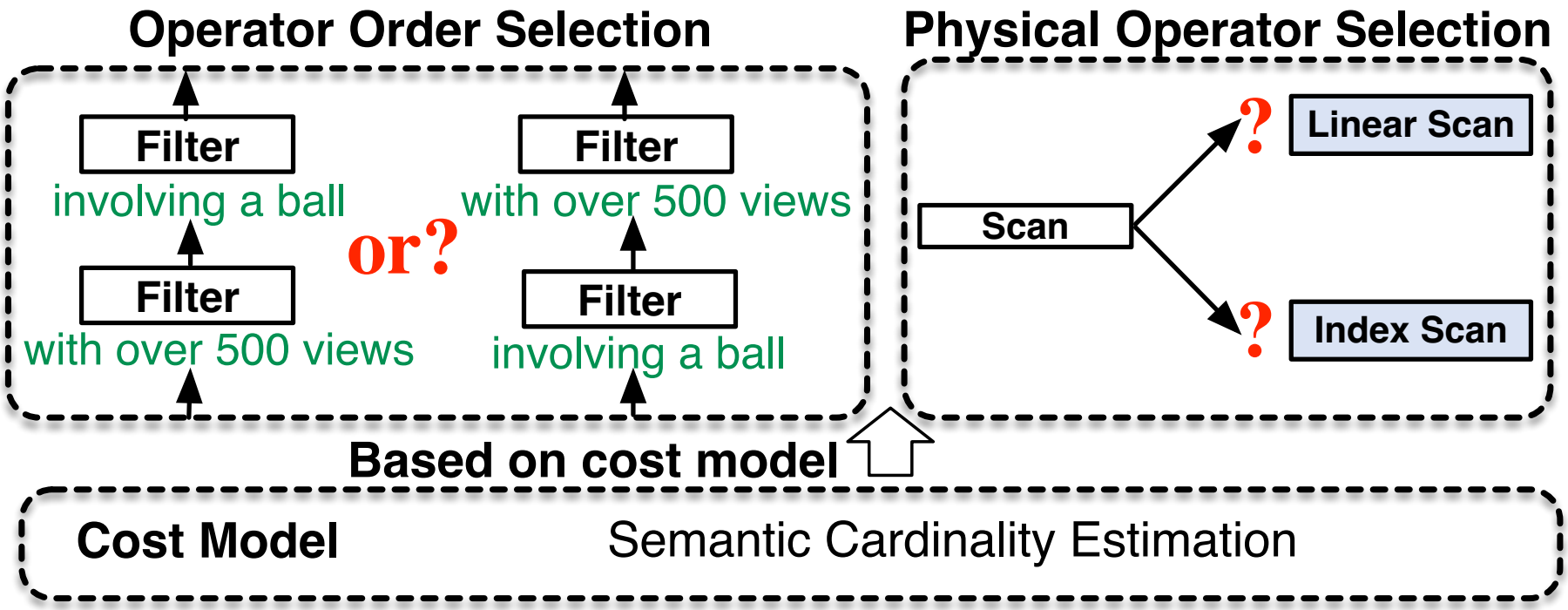


Pipeline Planning: Error Handling

- **Overview: progressively identifying appropriate pre-defined logical operators and reducing the query with the operators.**
  - ① **Semantic Parsing: extract the logical representations from the query**
  - ② **Operator Matching: identify the matched logical operators**
  - ③ **Query Reduction: reduce the logical operators to generate a plan**
  - ④ **Error Handling: backtrack to the previous reduction**
    - **If all candidate operators cannot reduce the query:**
      - **backtracks to the query before the latest reduction**
    - **If no reduction path can fully decompose the query :**
      - **Append a Generate operator to produce an answer based on collected information via LLM (RAG).**
      - **Instruct the LLM to generate Python code for solving the remaining task (fallback to code generation).**

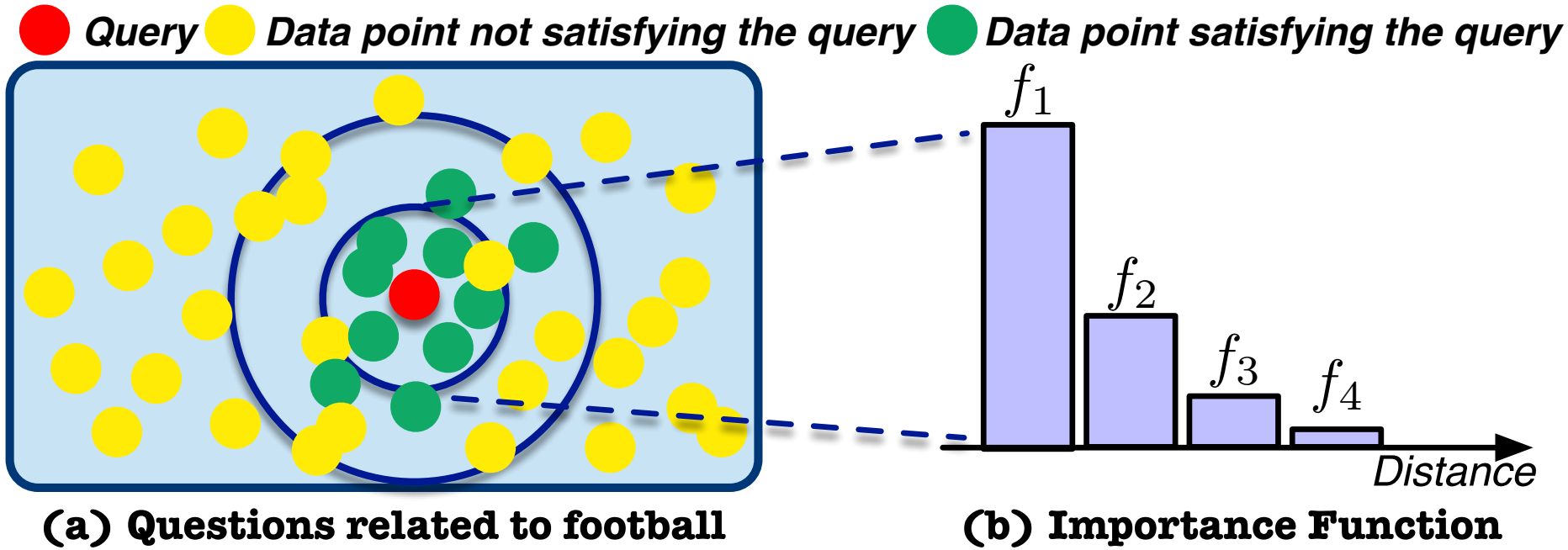
Pipeline Optimization

- Physical Operator Selection: LLM, Pre-Programmed, LLM Coding
- Cost Model:
  - LLM-based: Costs depend on input/output tokens and the number of LLM calls (Cardinality)
  - Program-based: Costs depend on input size and computational complexity



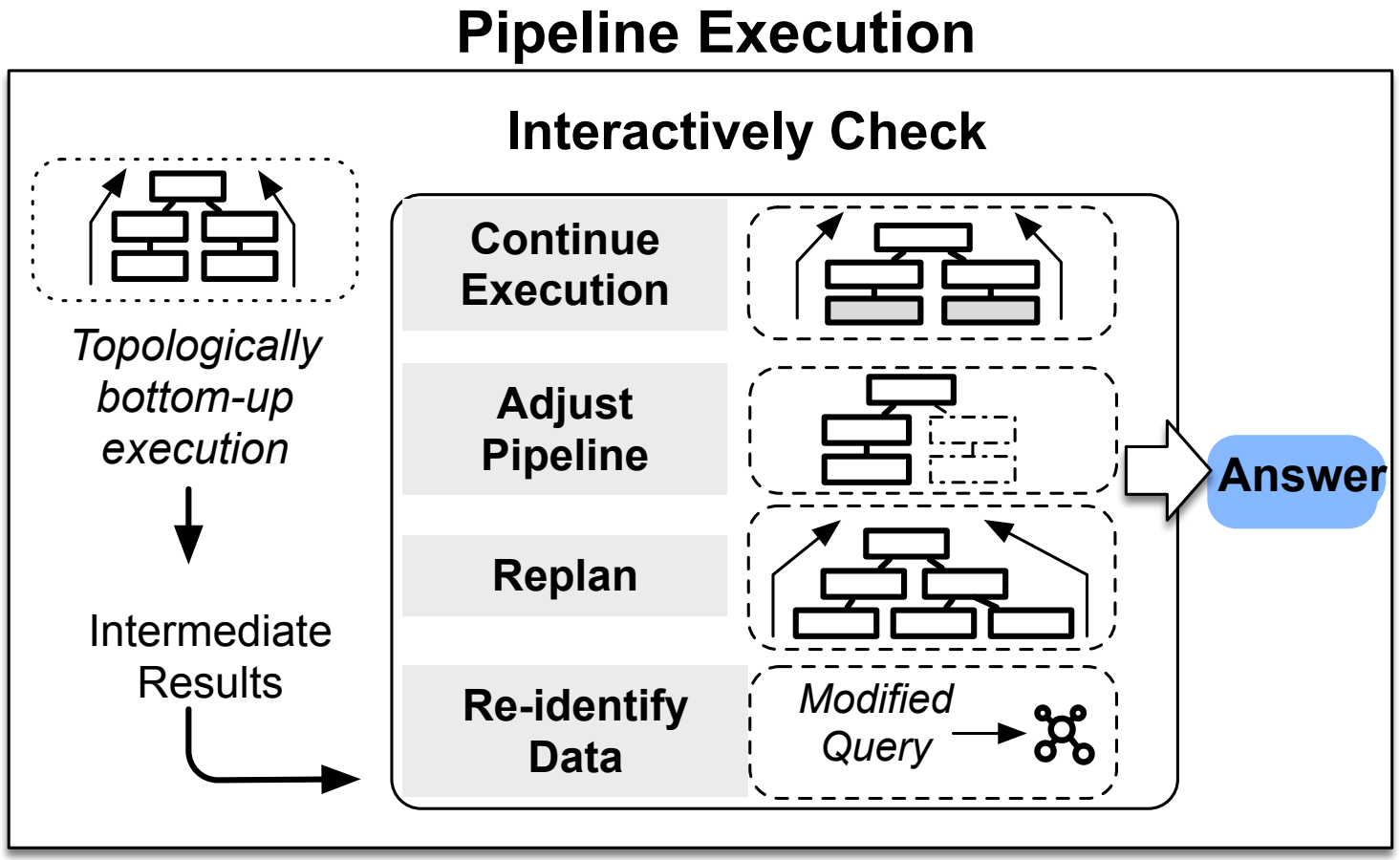
Pipeline Optimization: Semantic Cardinality Estimation

- Observation: data satisfying the query have small embedding distances to the query
- Key Ideas:
  - Estimation by importance sampling
  - Focus more on data points closer to the query vector



Pipeline Execution

- **Interactive Plan Adjustment During Execution:** When operator execution fails, dynamically adjusts the plan by continuously replanning
- Parallel Execution for low latency
- Multi-way Execution for high accuracy (Vector & LLM, Merge operators)

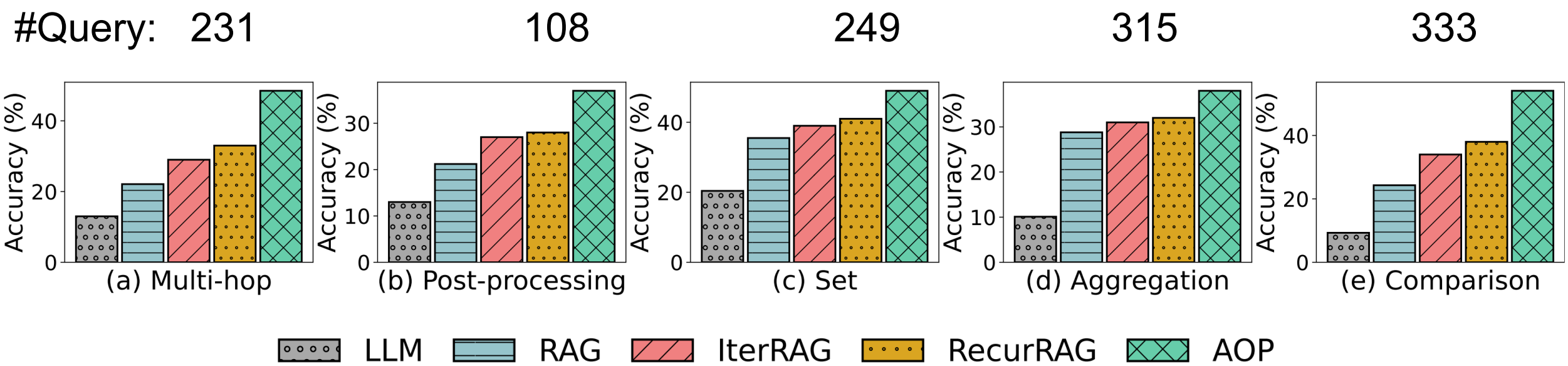


Pipeline Self-reflection

- Pipeline evaluation
  - Generate multiple pipelines, and select the best pipeline
  - Using samples to get some preliminary results and evaluating them
- Reward Model
  - Evaluate the utility of a pipeline and give a reward
    - Compute the utility of the current local pipeline
    - Predict the utility of the following global pipeline

Experimental Results

Real datasets and real NL queries

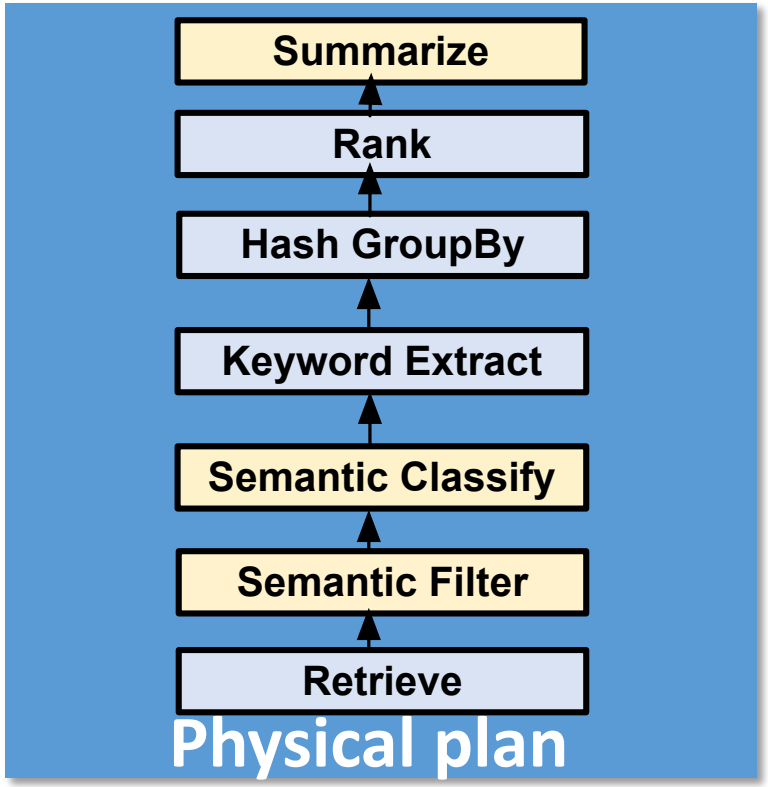


30%-40% accuracy improvement against GPT-4

Experiments: Case Study

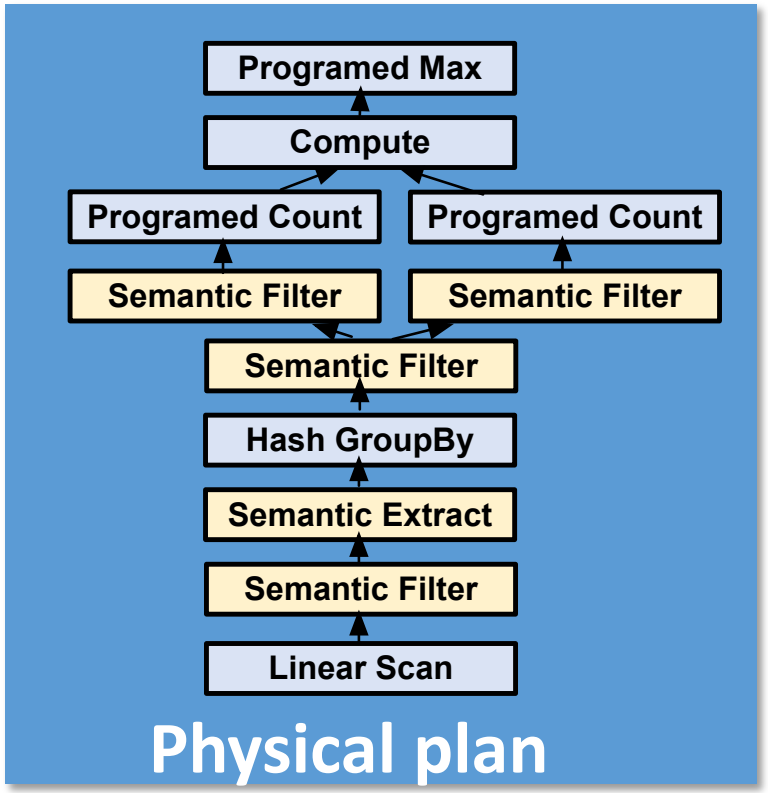
Which ball sport, among questions with over 500 views, has the highest ratio of injury-related questions to training-related questions?

Naïve



Fail to execute

Unify Agent



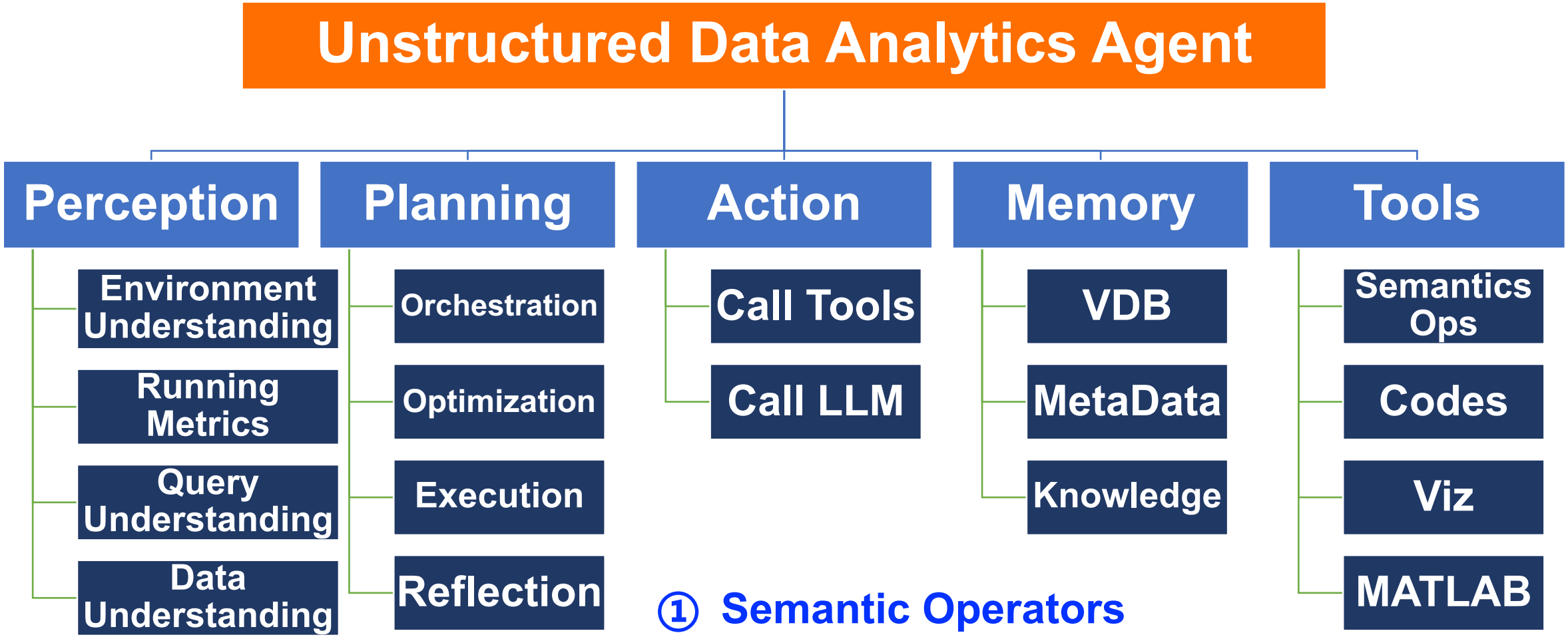
Obtain correct answer

Unstructured Data Analytics Agent

- ✓ Standard semantic operators for pipeline orchestration
- ✓ Automated pipeline orchestration for semantic data analytics
- ✓ Pipeline optimization & execution techniques
- ✓ Automatic support for multi-hop and semantic analytics
- ✓ Extensive experiments demonstrate that our method achieves both high accuracy and high efficiency.

Open source: https://github.com/TsinghuaDatabaseGroup/Unify

Unstructured Data Analytics Agent



① Semantic Operators
② Semantic Pipeline
③ Semantic Optimization & Execution

Jiayi Wang, Guoliang Li. AOP: Automated and Interactive LLM Pipeline Orchestration for Answering Complex Queries. CIDR 2025

Data Agent for Semantic Structured Data Analytics

Semantic Structured Data Analytics Agent

What are the number of positive and negative reviews for sci-fi movies given by users located in the capital of the US?

```
SELECT COUNT(Review)
FROM MovieReview
WHERE City =
SemanticExtract(US) AND
SemanticClassify(Movie)="sci-fi"
SemanticGroupBy(Review)
```

NL Queries

↓

NL2SQL

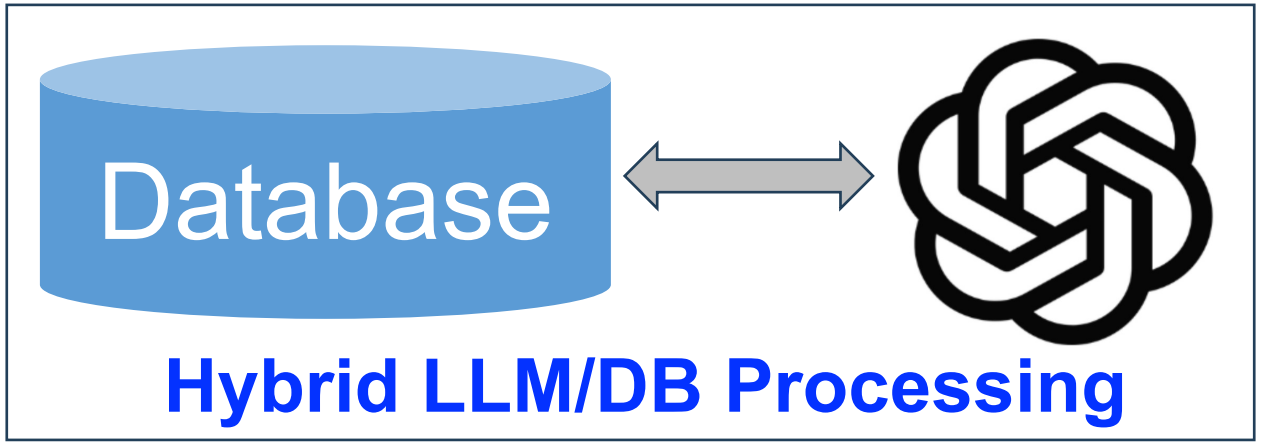
↓

Semantic SQL

↓

Semantic Filter
Semantic Groupby
Semantic Orderby
Semantic Classify
Semantic Join
Semantic Extract
Semantic Transform

Semantic SQL Processing



LLM As Semantic Operators

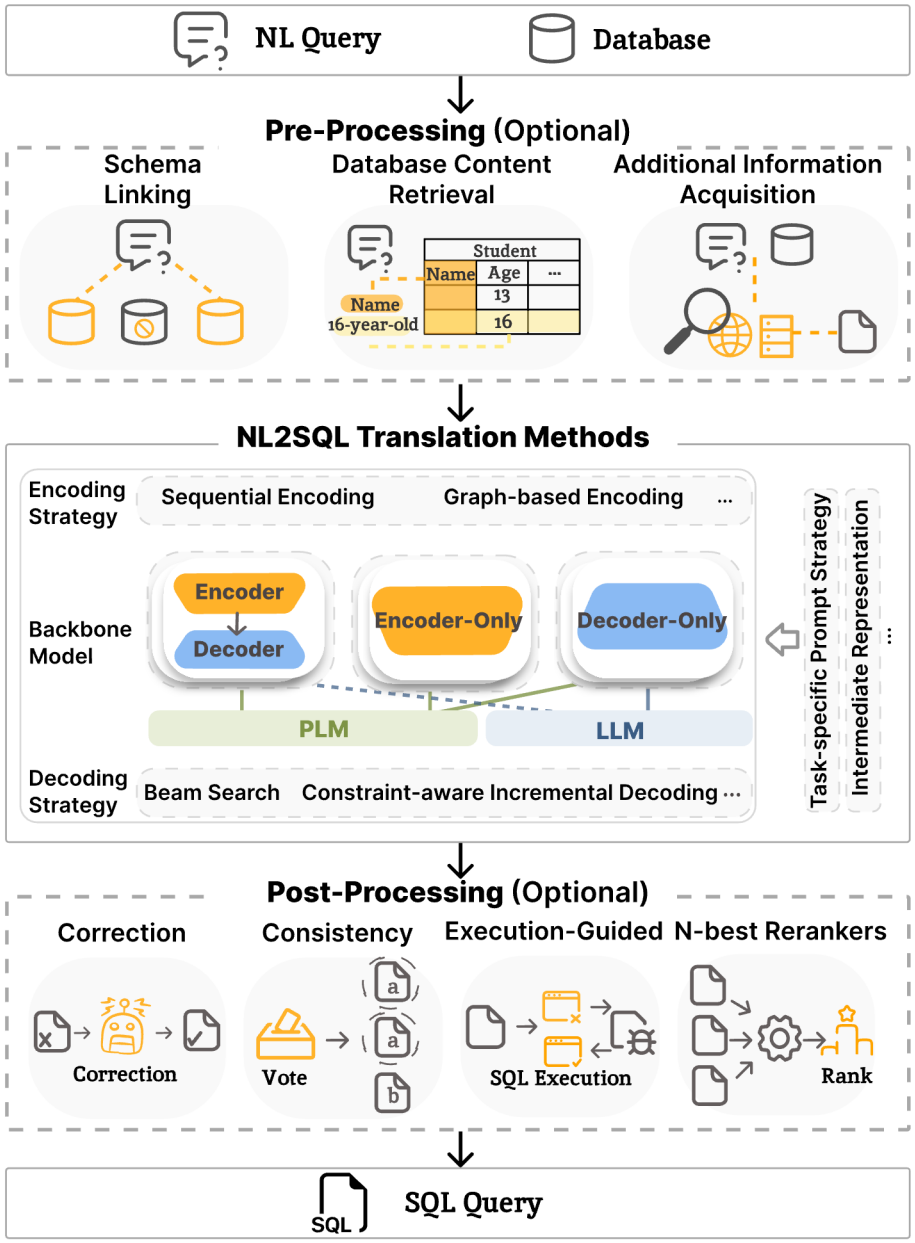
Semantic SQL: SQL with Semantic Operators

SQL + Semantic Operators: (LLM to execute Semantic operators)

- Semantic Acquire: Acquire external knowledge to answer a query
- Semantic Filter: Search for data entity semantically equal to query entity
- Semantic Classify: Categorize tuples based on semantic properties and criteria
- Semantic Group-by: Semantically group tuples
- Semantic Order-by: Semantically sort tuples
- Semantic Join: Semantically join tuples
- Semantic Transform: Transform SQL results to specific format or description

NL to Semantic SQL

- Query-Data Alignment
  - Query Rewrite
  - Query Augmentation
  - Schema Augmentation
  - Data Augmentation
  - Semantic Schema Linking
- NL2SemanticSQL Translation
  - Multiple SQL Generation
  - Best SQL Selection
  - SQL Reflection

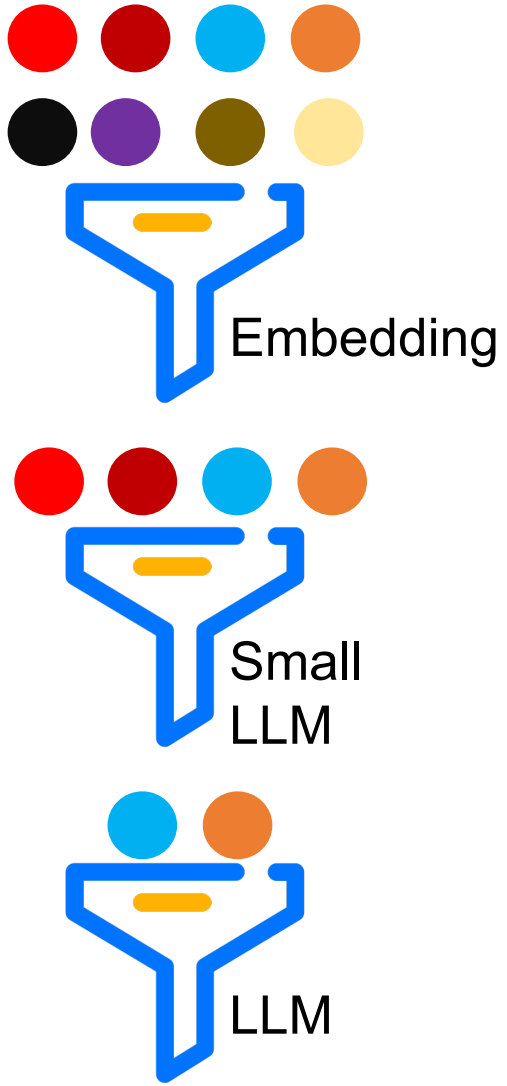


Semantic Operator Rewrite

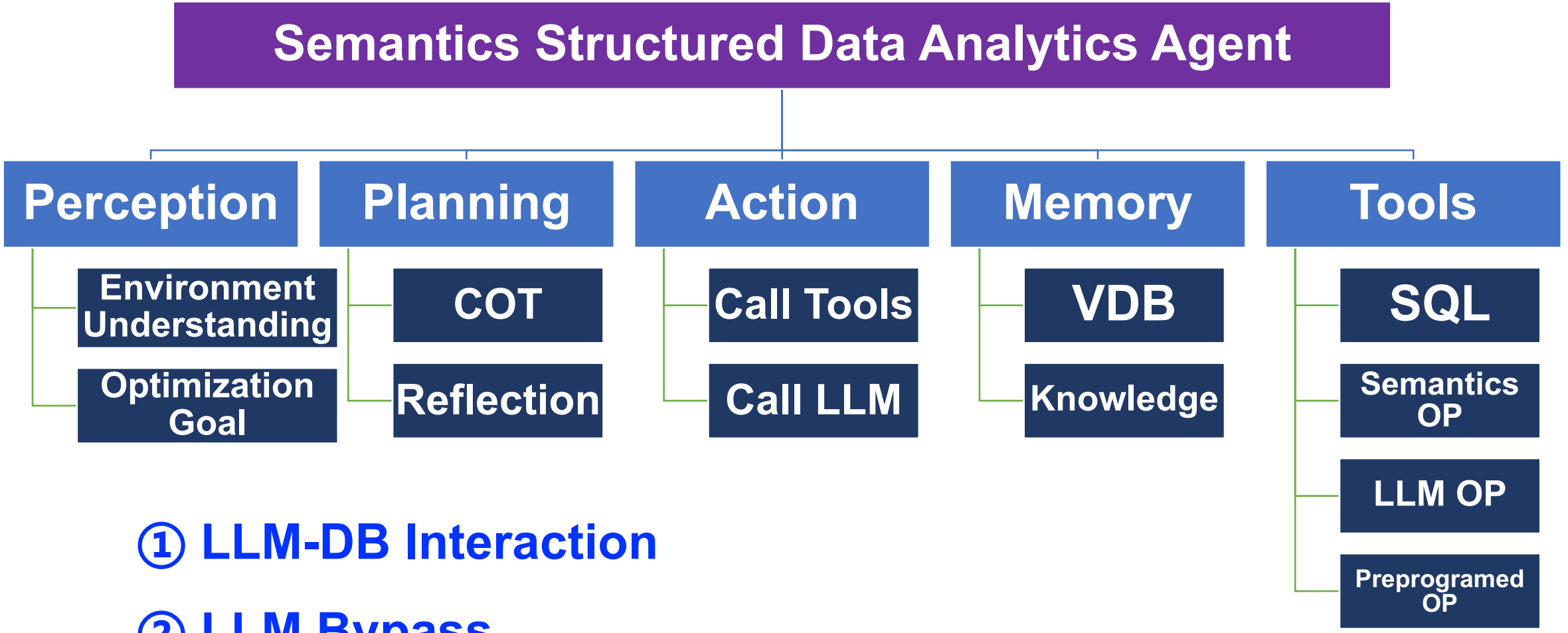
- Semantic operators are extensive.
- Reduce semantic operators to traditional operators
  - Semantic Acquire: report NBA players higher than Stephen Curry?
    ① Acquire the height of Stephen Curry from LLM
    ② Replace semantic acquire with height = 1.87m
  - Semantic Filter: report the conferences held in HK
    ① Identify cities that are semantically equivalent to HK, such as Hong Kong, Pearl of the Orient, Heung Gong.
    ② Replace semantic filter with “IN (HK, Hong Kong, Pearl of the Orient, Heung Gong)”

Semantic Plan Rewrite

- Invoking LLMs to execute semantic operators is expensive, it is vital to reduce LLM invocations
  - LLM Bypass
    - Using Embeddings/Vectors for Pruning
    - Using Small LLMs for Pruning
    - Using LLM Coding for Pruning (samples to verify the code)
  - Plan Rewrite
    - Push down: First DB operators and then LLM operators
    - Cost-based semantic operator order
  - Multi-Goal Optimization (Quality, Latency, Cost)



Semantic Structured Data Analytics Agent



① LLM-DB Interaction

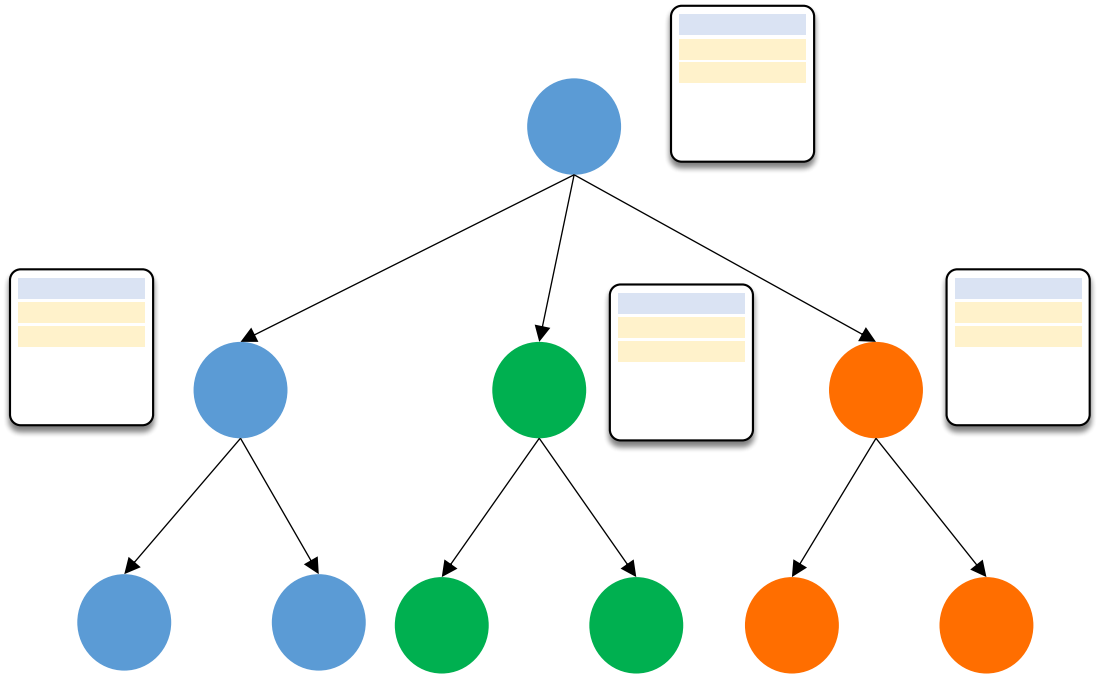
② LLM Bypass

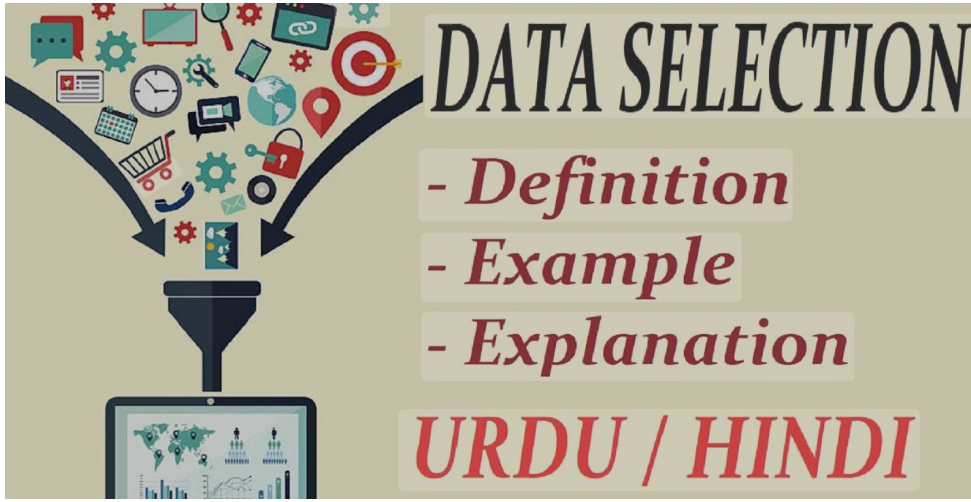
③ Semantic SQL Generation

Data Agent for Data Lake Analytics

Data Lake Analytics Agent: Challenges

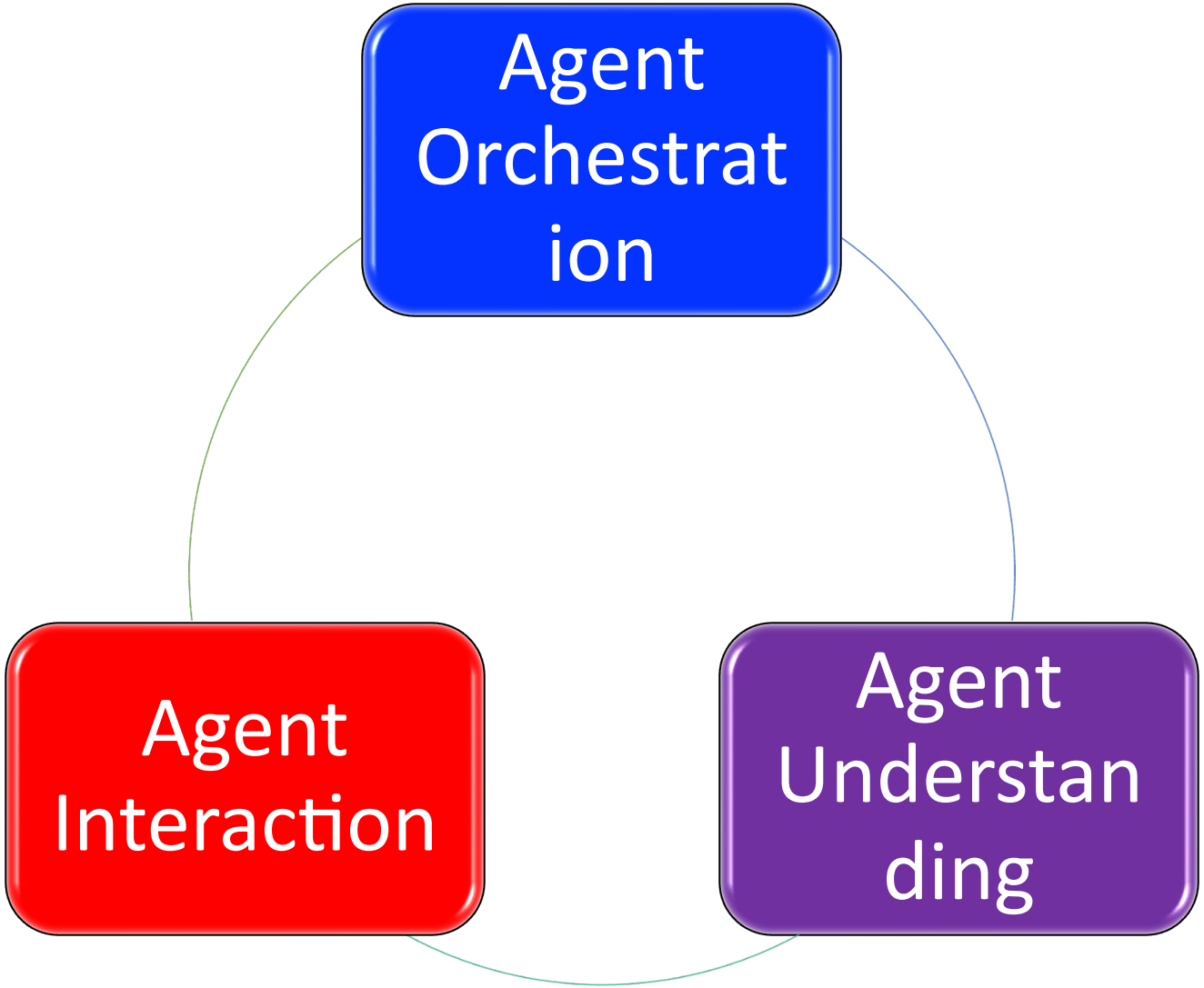
- Heterogenous Data
  - Data Linking
  - Data Embedding Model
  - OOD Indexing
- Semantic Source Selection
  - Unified Catalog
  - Semantic Data Organization
  - Hierarchical Data Exploration
- Pipeline Orchestration & Optimization
  - Agent Selection
  - Agent Interaction





Data Lake Analytics Agent: Multi-Agents

- **Pipeline Orchestrator Agent**
- **Data Agents**
  - **Active Meta Data Management**
  - **Hierarchical Data Selection**
  - **Semantic Catalog**
  - **Semantic Data Statistics**
- **Analytics Agents**
  - **Unstructured Data Analytics**
  - **Structured Data Analytics**
- **Tool Agents**

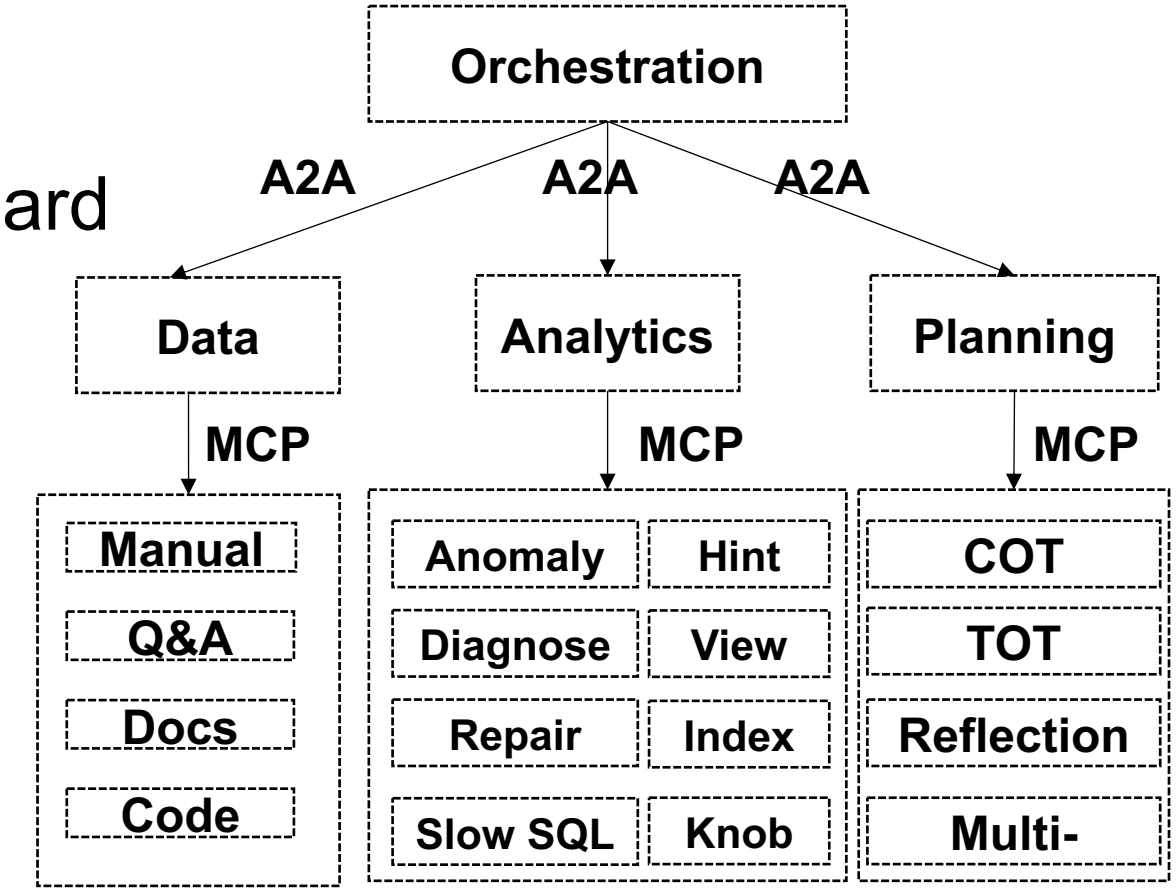


Why Multiple Agents

- **Enhanced Data Analytics:**
  - **Unstructured Data Analytics Agent**
    - **ETL-based, Specific Data Types, Orchestration-based**
  - **Structured Data Agents**
- **Modularity**
  - **Data Analytics, Data Visualization, DB Interaction**
- **Collaboration and Coordination: leveraging shared knowledge and resources.**
- **Robustness and Fault Tolerance: the failure of one does not necessarily compromise the entire system**
- **Parallelism and Efficiency**

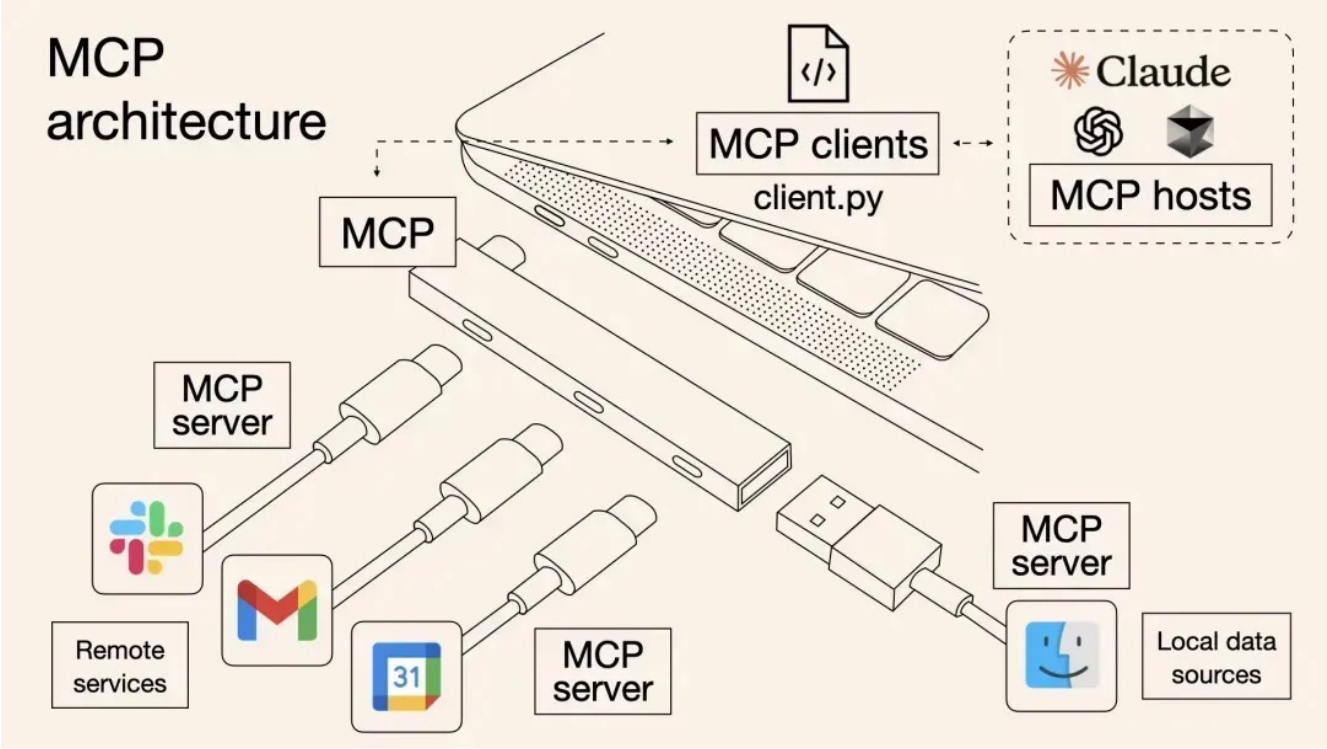
Agent-Agent Interaction Management

- Agent: Control Plane
- Agent Capability Discovery → Agent Card
  - Agent Role Understanding
  - Agent Role Description and Augmentation
  - Agent authentication
- Agent Communication – A2A Protocol
  - Agent message
  - Agent status & lifecycle – submitted, ready, working, completed



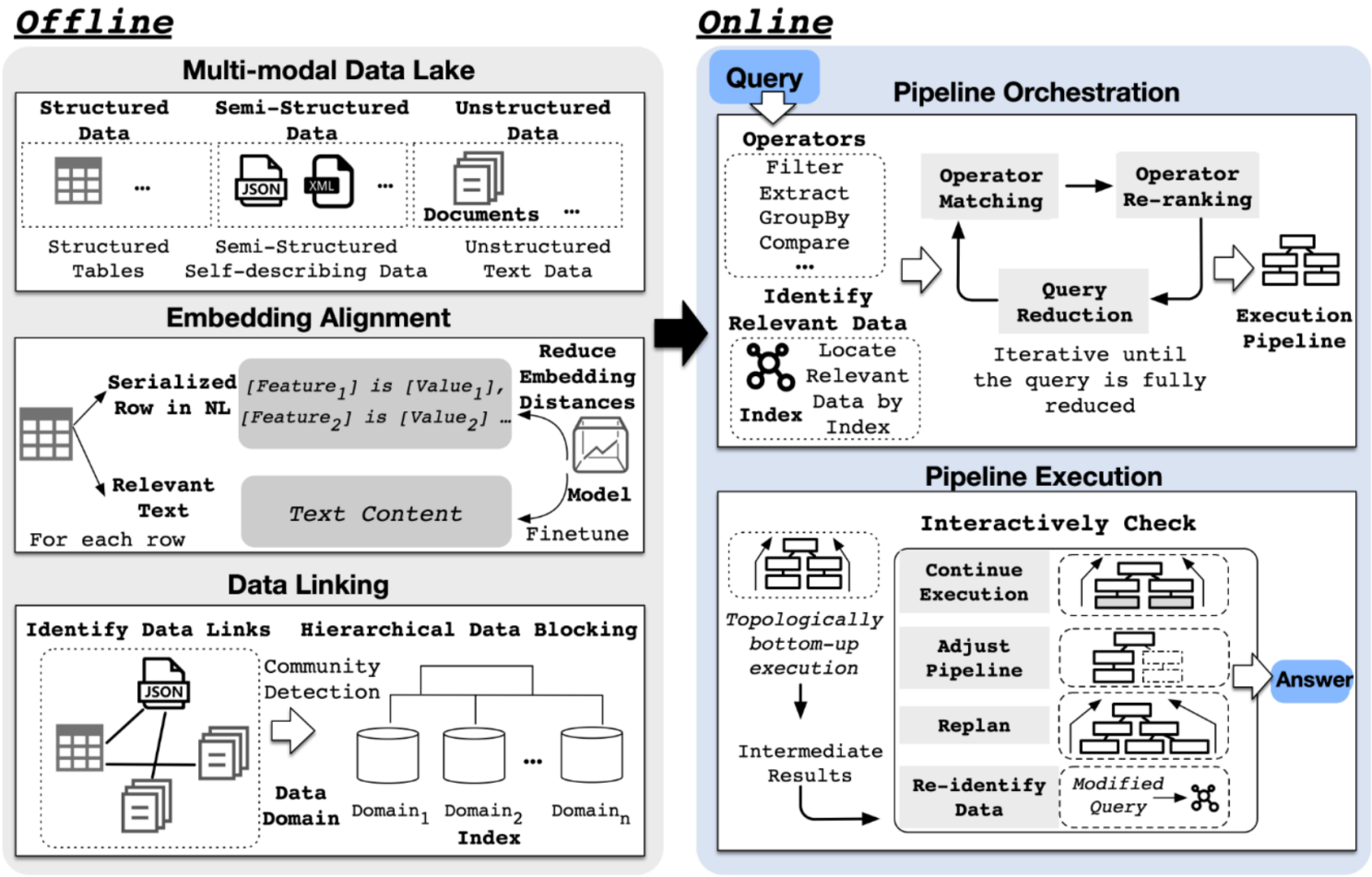
Agents-Tools Interaction

- **Model Context Protocol (MCP)**
- **Input Alignment:** All models receive consistent and traceable information, eliminating the need for each to make independent guesses.
- **State Synchronization:** During task execution, state changes are recorded and broadcast in a standard format to prevent information drift.
- **Inference Consistency:** Intermediate inference results from different models can be read and reused by each other, forming a coherent chain of reasoning.



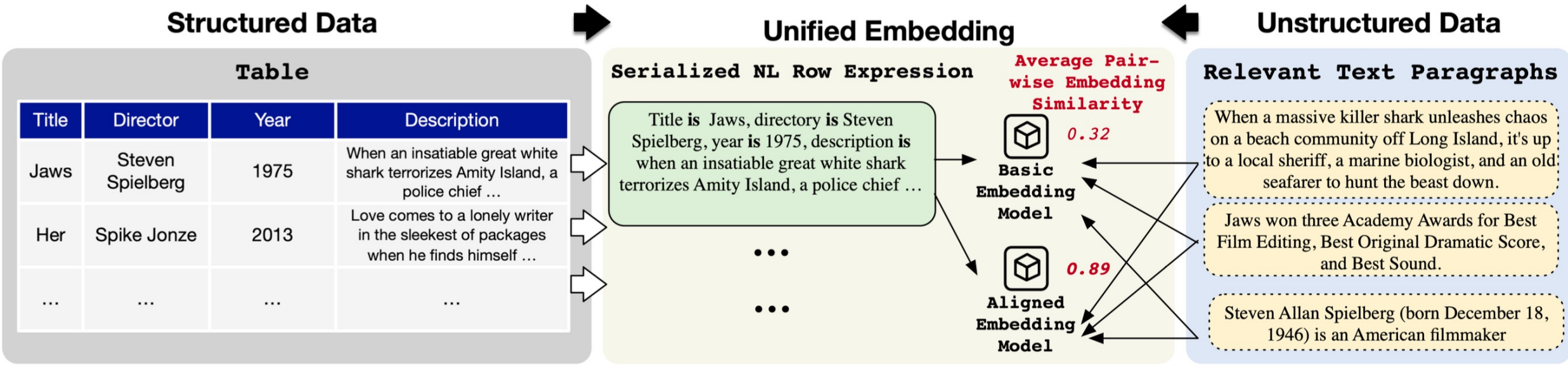
Data Lake Analytics Agent : Architecture

- Offline Data/Knowledge Preparation
  - Data Preparation for RAG and Prompt Engineering
    - Semantic Segmentation
    - Vector Indexing
    - Data Linking
  - Environment/Tool Understanding
    - Fine-tuning for alignment
    - Tool Calling
- Online Agent for optimization
  - Pipeline Orchestration
  - Pipeline Execution
  - Agent Interaction



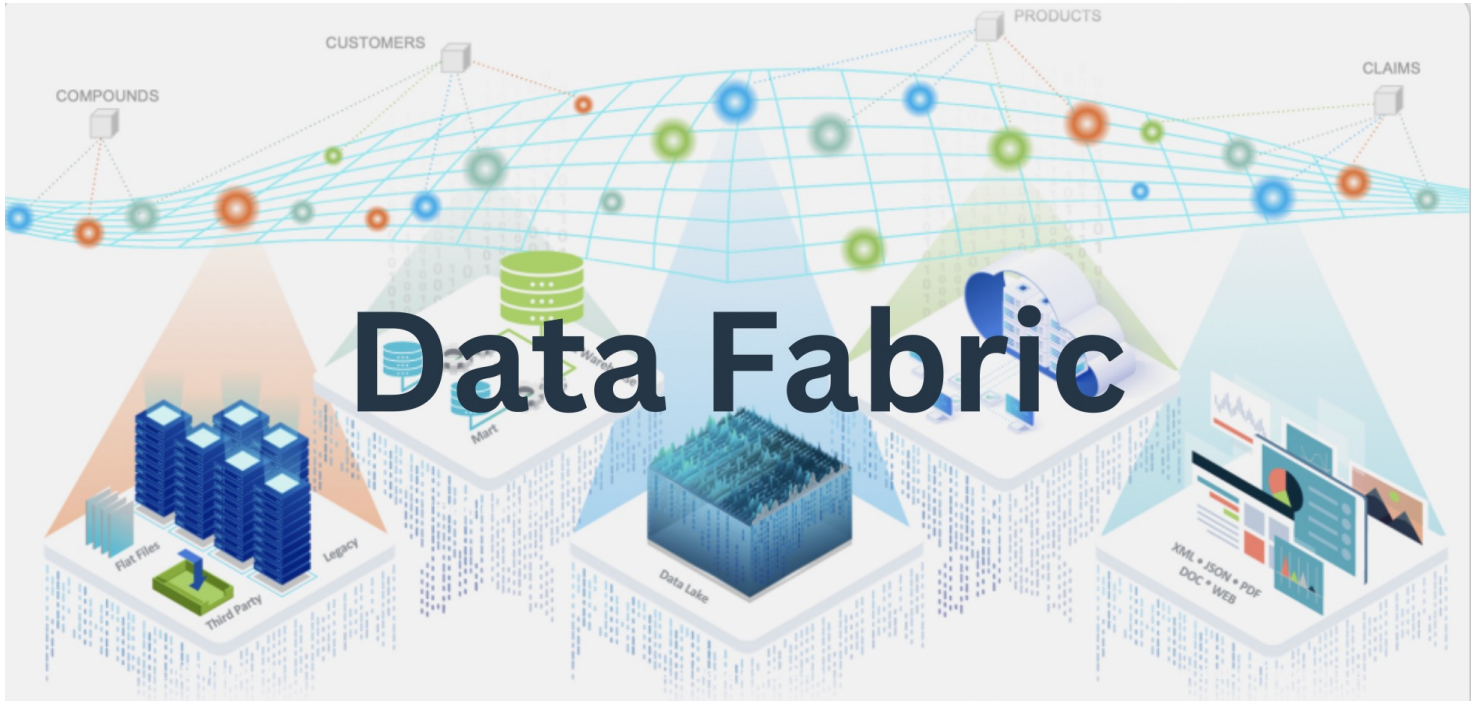
Data Lake Analytics Agent: Data Fabric

- Unified Embedding
  - A bi-encoder-based method that transforms diverse data types into a unified embedding space
  - Data annotation and augmentation



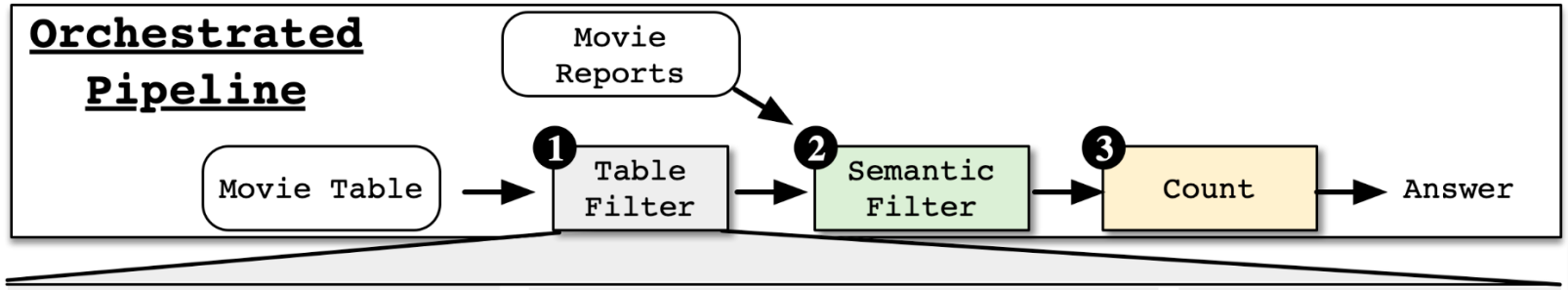
Data Lake Analytics Agent: Data Fabric

- Unified Data Access: Provides a single, consistent interface for accessing data, facilitates real-time data access and sharing across the organization.
- Semantic Catalog and Semantic Data Organization
- Active Meta Data Management and Update
- Automatic Data Pipelines
- Data Lineage and Provenance
- Support for Diverse Tools
- Self-Service Analytics



Data Lake Analytics Agent: Orchestration

- Pipeline Orchestration: Iterative Query Decomposition
  - Coarse-grained matching: identifying suitable agents
  - Fine-grained re-ranking: identifying suitable operators in agents
  - Reducing: applying a feasible operator to simplify the query



Operator Matching

Semantic Parsed Query

Count the total number of [Entity] about [Entity] [Condition] in the Movie Table

All Operators	Semantic Similarity
Semantic Filter	0.94
Table Filter	0.92
Extract	0.34
GroupBy	0.29
...	...

Operator Re-ranking

Candidate Operators → Verify → Re-ranked Operators

- Semantic Filter
- Table Filter

✓ Table Filter
✗ Semantic Filter

In the verification

- Check the logical feasibility
- Check the input availability
- Check operator dependency

Query Reduction

Q1 → Rewrite → Q2

Example Prompt

Given the query Q1 and a matched logical presentation, please rewrite query Q1 by reducing the matched segment of the logical representation

Query reduction during the process

Input query Q1 for 1

Count the total number of reports about movies directed by Steven Spielberg in the Movie Table

Input query Q2 for 2

Count the total number of reports about the movies.

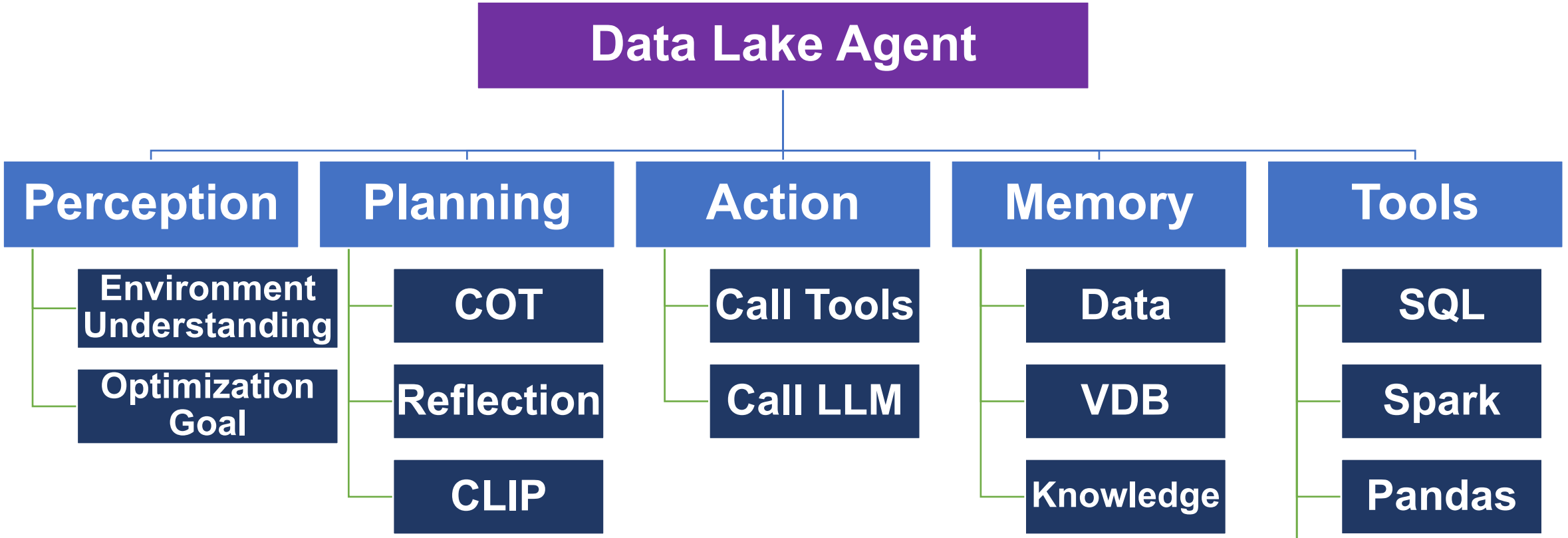
Input query Q3 for 3

Count the total number of reports

Completely reduced result

Number

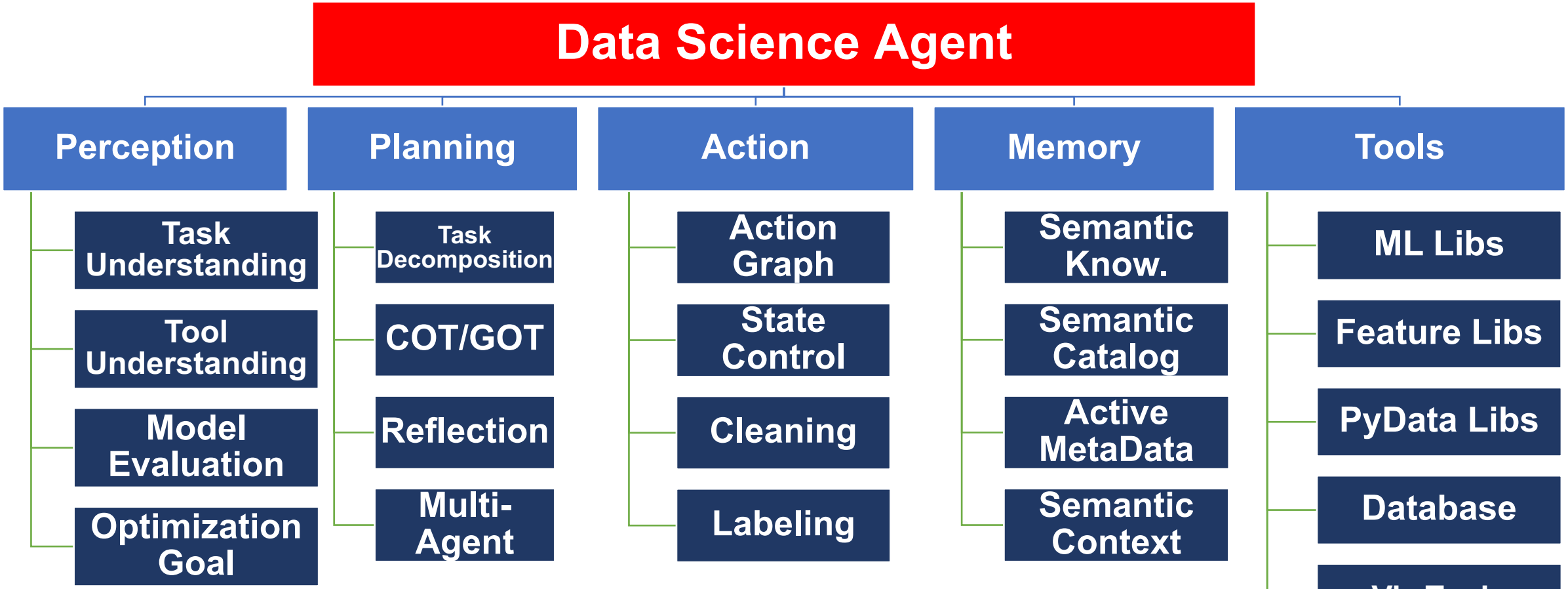
Data Lake Analytics Agent



① Data Linking
② Semantic Catalog & Active Meta Data
③ Agent Selection & Interaction

Jiayi Wang, Guoliang Li, Jianhua Feng: iDataLake: An LLM-Powered Analytics System on Data Lakes. IEEE Data Eng. Bull. 49(1): 57-69 (2025)

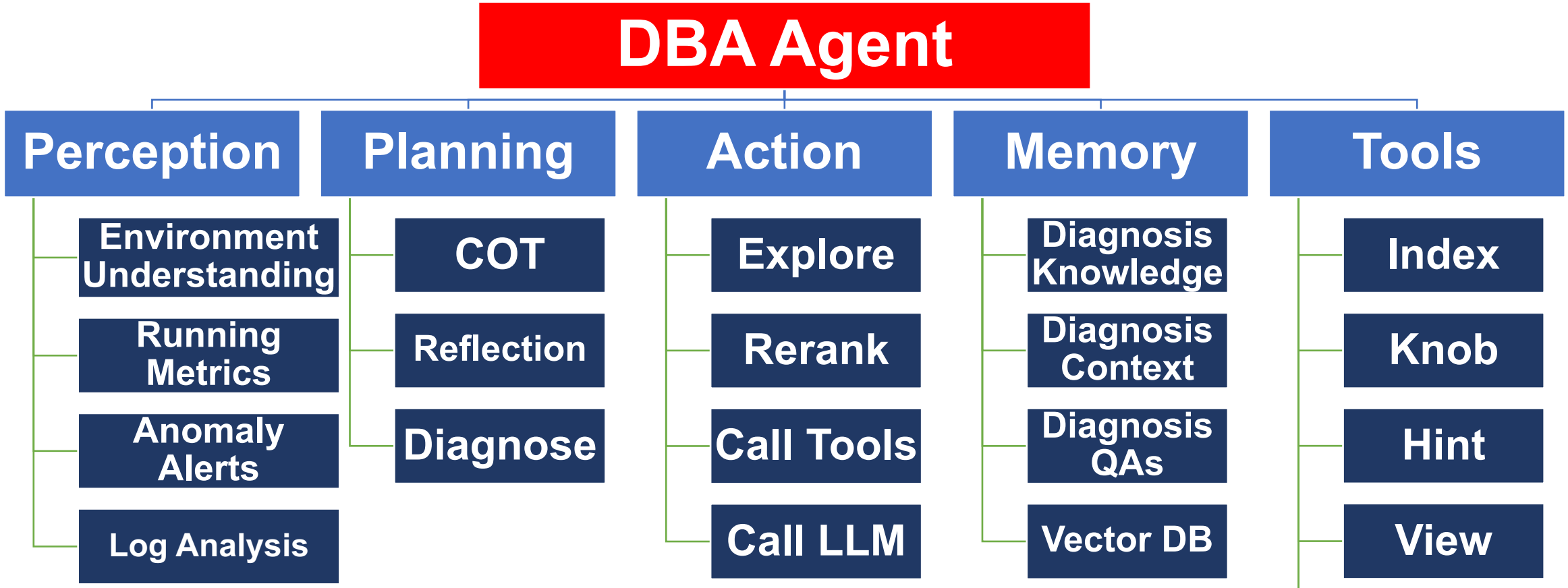
Data Science Agent



① Data Science Data & Tool Understanding
② Data Science Task Understanding & Decomposition
③ Data Science Pipeline Reasoning & Planning
④ Multiple Agents Collaboration & Interaction

Yihang Zheng et al. Revolutionizing Database QA with Large Language Models: Comprehensive Benchmark and Evaluation. KDD 2025

Database Administrator Agent



Preparation & Summarization
Knowledge Extraction
Tool Learning
Environment Learning

Tree Search Based Diagnosis
Self-Reflection
Collaborative Multi-Agent
Localized fine-tuning

Xuanhe Zhou, Guoliang Li, et al: D-Bot: Database Diagnosis System using Large Language Models. VLDB. 17(10): 2514-2527 (2024)

Data Agent Opportunities in Data Lifecycle Management

① Data Generation → ② Data Transform → ③ Data Analytics → ④ Data Engineering → ⑤Data Usage → ⑥Data Interpretation

① OLTP

- Regression
  - Cardinality/Cost Estimation
- Online NP Optimization
  - Query Rewrite
  - Plan Selection
- Offline NP Optimization
  - Knob/Index/View Advisor
- Prediction
  - Workload
  - Resource
  - Data
- Database QAs
- Database Diagnosis
- SQL Dialect Translation

② Data Transformation

- Data Standardization
  - Automate Extract, Transform, Load
  - Auto schema mapping
  - Identifying patterns
  - Low code ETL
  - Predictive auto-scaling
  - Adaptive Transformation
- Multi-goal Optimization
  - Change data capture
  - Automate CDC
  - Predictive analytics
  - Reduce Human Cost

③ OLAP

- In-Database ML
  - In-Database Model
  - In-Database Vector
  - In-Database RAG
  - Anomaly Detection
  - Risk control analysis
- SQL+ML Analytics
  - TableQA
  - In-DB Semantics Analytics
- Autonomous Analytics
  - Autonomous workload management
  - Autonomous data format
  - Autonomous serverless

Data Agent Opportunities in Data Lifecycle Management

① Data Generation → ② Data Transform → ③ Data Analytics → ④ Data Engineering → ⑤Data Usage → ⑥Data Interpretation

④ Data Engineering

- Data Preparation
  - Data Discovery
  - Data Selection
  - Data Cleaning
  - Data Transformation
  - Data Integration
  - Data Generation
  - Data Mixing
  - Data Extraction
  - Data Labeling
  - Meta Data Manag.
- Data Flywheel
- Data Fabric

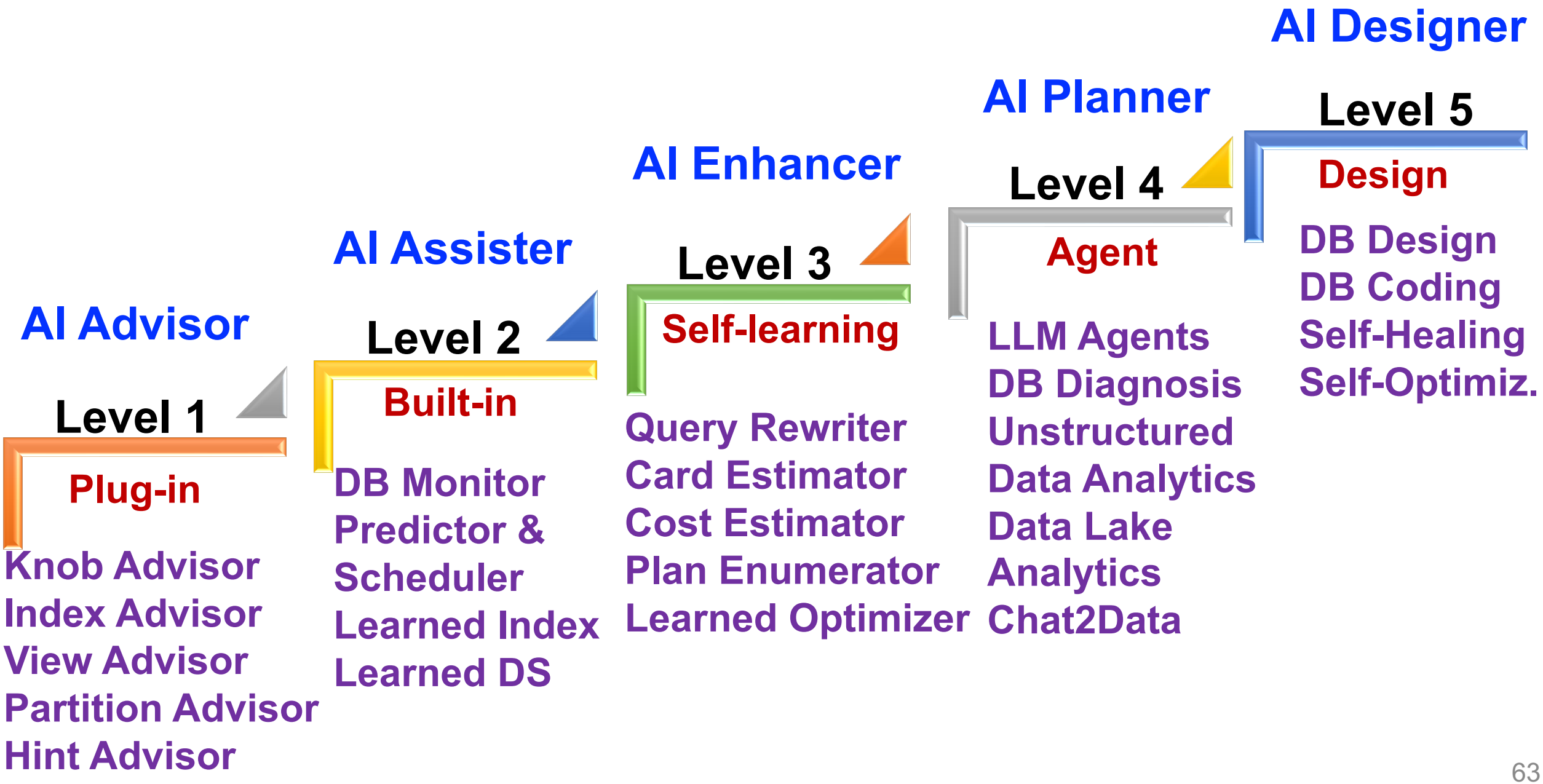
⑤ Descriptive Data Usage

- Prompting
  - Automation/Examples
- RAG
  - Multi-hop RAG
  - Graph RAG
  - Agentic RAG
- LLM Inference
  - Prefill/Decoder disagg.
  - KV cache
  - Scheduling
  - Data/Model/Resources Parallism
  - Quantization

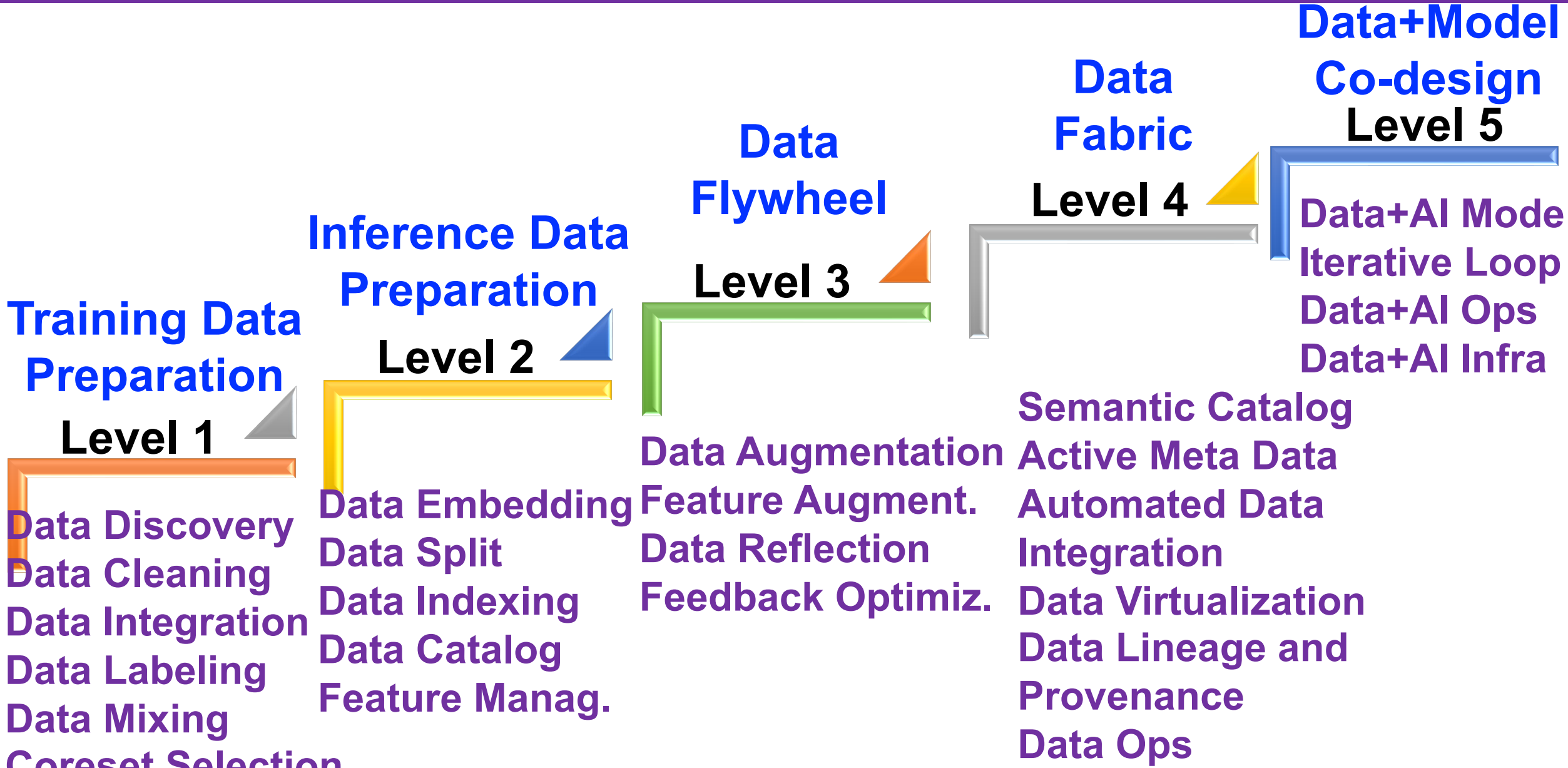
⑥ Proactive Data Interpretation

- Proactive Insights
  - Trends summarization
  - Insight discovery
  - Predictive decision making
  - Cognitive Analytics
- Proactive Data Agent
  - Chat2Data
  - Chat2BI
  - NL2Viz
  - In-DB Semantics Analytics
  - Unstructured data analytics
  - Multi-model data analytics on data lake

Five Levels of AI4Data



Five Levels of Data4AI



Conclusion

- Data+AI is important for data management and analytics
- It urges the use of Data+AI techniques to revolutionize data systems
  - Data science, Data Analytics, Data Lake
- Data Agent is a promising direction for Data+AI
  - Agent Orchestration and Scheduling
  - Multi-Agents Interaction
  - Tool Calling
  - Proactive Data Management
- Open-source systems for Data Agent

Thanks!

Slides: *https://dbgroup.cs.tsinghua.edu.cn/ligl/activities.html*

Data+AI Paper List: *https://github.com/code4DB/LLM4DB*

System: *https://github.com/TsinghuaDatabaseGroup/Unify*