

 Enterprise Data AI Strategy

Enterprise Data Analytics Agents

Architectural Deep-Dive: Common Foundations, Context Construction Philosophies, and
Databricks Genie's Living Ontology

Databricks Genie

Snowflake Cortex

Google Looker / BigQuery

1. 기업 데이터 분석용 Agents

기업 데이터 분석용 Agent는 대화형 AI 에이전트를 통해 기업 데이터 접근 문제를 해결하는 것을 목표로 함

Platform	User-Facing System	Internal Analytical Components
 Databricks	Genie One · Genie Agents	Genie Ontology, Agent Mode, Databricks SQL Engine
 Snowflake	Cortex Agents	Cortex Analyst (Text-to-SQL), Cortex Search, Semantic Views
 Google Cloud	Looker / BigQuery Conversational AI	Gemini, LookML Semantic Model, BigQuery Analytics Engine

<https://www.databricks.com/kr/blog/introducing-genie-one-genie-ontology-and-genie-agents>

2. 기업용 분석 Agent Workflow

Text-to-SQL은 일부일 뿐, 핵심은 governed analysis loop



① 자연어 Query

“최근 4주간 제품 P7의 수율이 하락한 원인을 장비, Recipe 변화, Lot 구성 관점에서 분석해줘. Engineering wafer는 제외하고 공식 수율 기준을 사용해줘.”



② Context Grounding

SELECT 검증 SQL
e.equipment_group,
production_week,
SUM(w.good_die) / SUM(w.tested_die)
AS yield ...

공식 수율 = SUM(good) / SUM(tested)

“Engineering wafer”

→ lot_class_cd = 'ENG'

Semantic layer

‘최근’은 완료된 production week 기준이다.



③ 실행 경로 설정

SELECT
e.equipment_group,
production_week,
SUM(w.good_die) / SUM(w.tested_die)
AS yield ...

공식 수율 = SUM(good) / SUM(tested)

“Engineering wafer”

→ lot_class_cd = 'ENG'

‘최근’은 완료된 production week 기준이다.



④ SQL 생성 (text-to-SQL)

SELECT
...
WHERE w.product_id = 'P7'
AND l.lot_type <>
'ENGINEERING'
AND production_week
BETWEEN '2026-W25' AND
'2026-W28'
GROUP BY
e.equipment_group,
production_week;
...



⑤ 반복 실행 및 Agent 분석

SQL 1: 주간 수율 추세

SQL 2: 장비별 수율

SQL 3: Recipe별 수율

SQL 4: Recipe 변경 전후 비교

SQL 5: Lot mix 분석

3. 플랫폼별 Context 선택 전략

Databricks Genie

관련 개념·SQL·Dashboard
탐색

출처 권위·사용량·최신성 평가

높은 authority context로
SQL 실행

"조직에서 실제 어떤 지표와 쿼리가 가장 권위
있게 자주 쓰였는가?"

Snowflake Cortex

Semantic View /
Model 확인

Verified Query
Repository 탐색

승인된 정의·SQL을 재사용

"데이터 전문가가 어떤 정의와 SQL을
공식 승인(Verify)했는가?"

Google Looker/BigQuery

LookML 또는 Data Agent
context 확인

field·measure·filter에 매핑

semantic model 규칙으로
query 구성

"명시된 semantic model이 무엇인가"

📄 전략 요약: Databricks는 Authority 추론, Snowflake는 승인 정의 재사용, Google은 코드 명시 실행 전략을 취함.

4. 기존 Explicit Semantics의 문제점

1. 지표 변경시 데이터 엔지니어가 Context들을 일일이 수작업 수정해야 함.
2. 기업의 “공식 의미”는 하나의 문서에만 있지 않아, Context가 흩어져 있다

Manual Curation의 한계

기존의 정적인 Semantic Layer는 모든 지표, 테이블 간 JOIN, 동의어, 업무 규칙을 데이터 엔지니어가 직접 정의하는 데 크게 의존함.

⚠ Manual Curation 비용 때문에, 기업의 Business Logic이 빠르게 변화할수록 기존의 Manual Curation과의 차이가 발생한다..

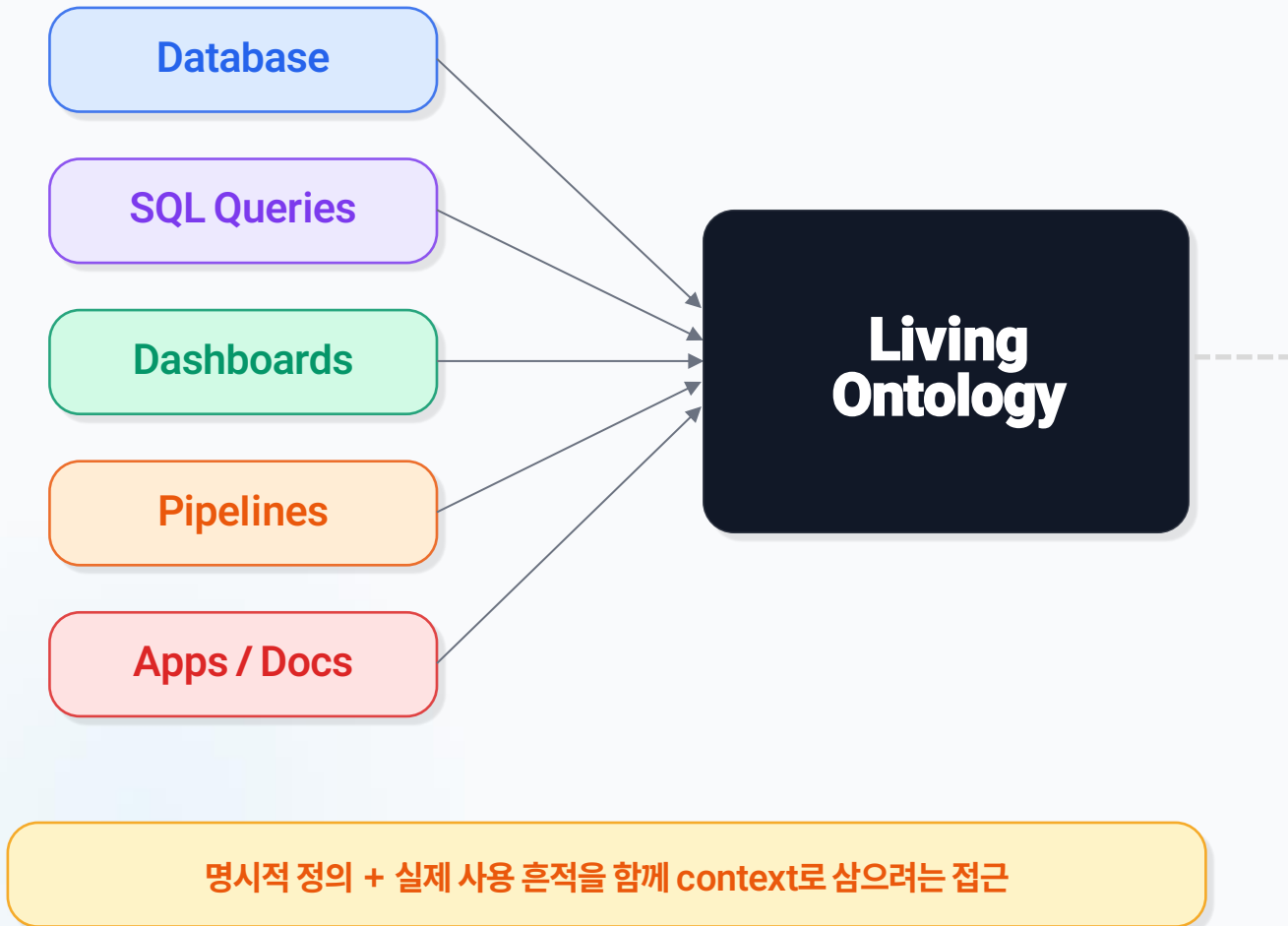
분산되어 있는 기업 Knowledge

실제 기업의 business definition은 하나의 Modeling file에만 존재하는 경우가 거의 없음. 오래된 SQL 스크립트, BI 대시보드, ETL 파이프라인, Teams 채팅 등 여러 곳에 분산되어 있음.

💡 Databricks Genie는 워크스페이스 전반의 운영 자산을 지속적으로 탐색해 이러한 문제를 해결하고자 함

5. Databricks Genie : Living Ontology Graph

Implicit Semantic의 생성

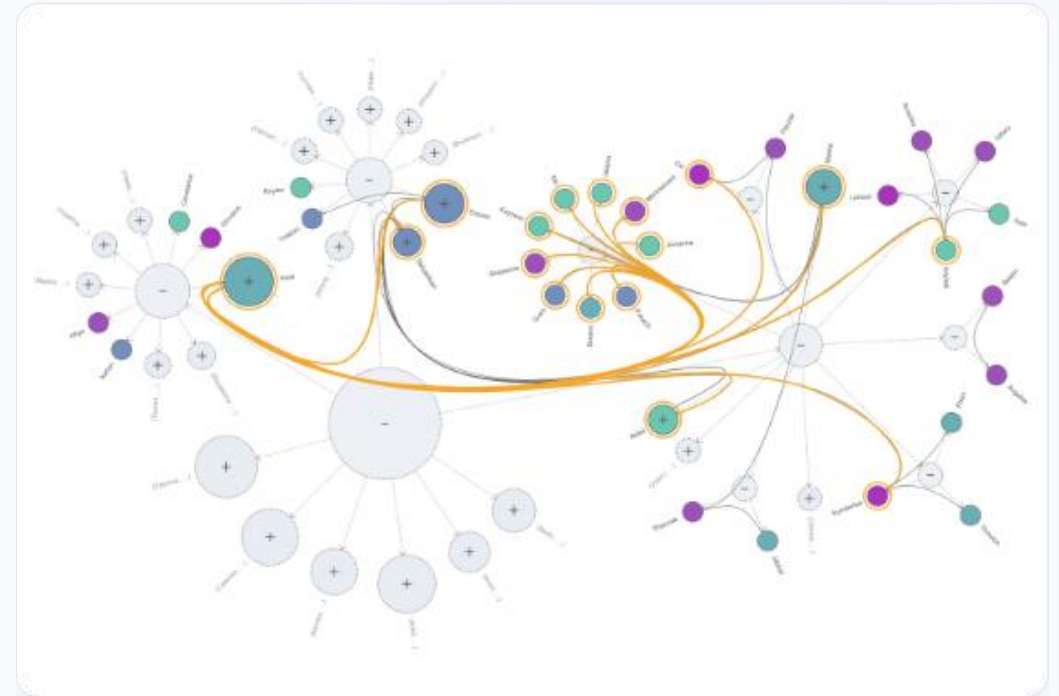


Living Ontology Graph

$$G = V, E$$

*Node (V): 수율, 매출, 장비, 테이블, SQL, 대시보드, 파이프라인, Teams

*Edge (E): 계산 규칙, 조회 관계, 데이터 생성 관계, 관리 권한 관계



6. Databricks Genie : Authority-Aware Ranking

충돌하는 정답의 순위화: OntoRank Scoring Model

$$\text{Authority (s)} \approx w_1 \text{SourceTrust} + w_2 \text{Usage} + w_3 \text{Certification} + w_4 \text{Freshness} + w_5 \text{GraphCentrality}$$

※ 개념적 표현. 실제 scoring function과 weight는 공개되지 않음.

Source Trust

공식 파이프라인 vs 임시 쿼리 출처 신뢰도

Usage Frequency

조직 내에서 얼마나 자주 실행되고 호출되었는가

Certification

데이터팀이 공식 인증 자산으로 지정했는가

Freshness

최근에 업데이트되고 정상 관리된 지표인가

Graph Centrality

핵심 대시보드 및 지표들과 밀접하게 연결되었는가

Source Trust & Certification

Unity Catalog에서 certified assets, 검증된 테이블, 그리고 도메인 전문가가 작성한 내용을 우선적으로 평가

Usage & Freshness

자주 실행되는 쿼리, active BI 대시보드, 최근에 업데이트된 파이프라인에 더 높은 가중치를 부여

 Cold Start 보완책: 사용 이력이 없는 신규 데이터는 엔지니어가 등록한 Trusted Logic / Metric View가 최우선 가중치를 가짐.

7. Living Ontology Graph vs Explicit Control

Living Ontology Graph의 Philosophy

Databricks Genie: 워크스페이스에서 계속 생성·변화하는 자산으로부터 Context를 자동으로 발견하는 데 초점을 둠

- + 유지·관리 부담이 비교적 낮음, 조직의 변화에 동적으로 적응
- 자주 사용되는 오래된 쿼리가 새롭고 올바른 로직보다 일시적으로 더 높은 우선순위를 차지할 수 있음

Explicit Control의 Philosophy

Snowflake & Google: 사람이 명시적으로 정의하고 검토한 내용을 최종적인 기준, 즉 Source of Truth로 삼는 방식을 강조

- + 감사와 추적이 쉬움. 결과가 더 deterministic 하고 정밀하게 통제됨
- 관리 작업이 많이 필요함. 데이터팀이 정의와 모델을 유지하는 과정에서 병목이 발생할 수 있음

⚠ Authority ≠ Statistical truth