제조고객을 위한 Industry Insight Day

Beyond GenAI & Demo

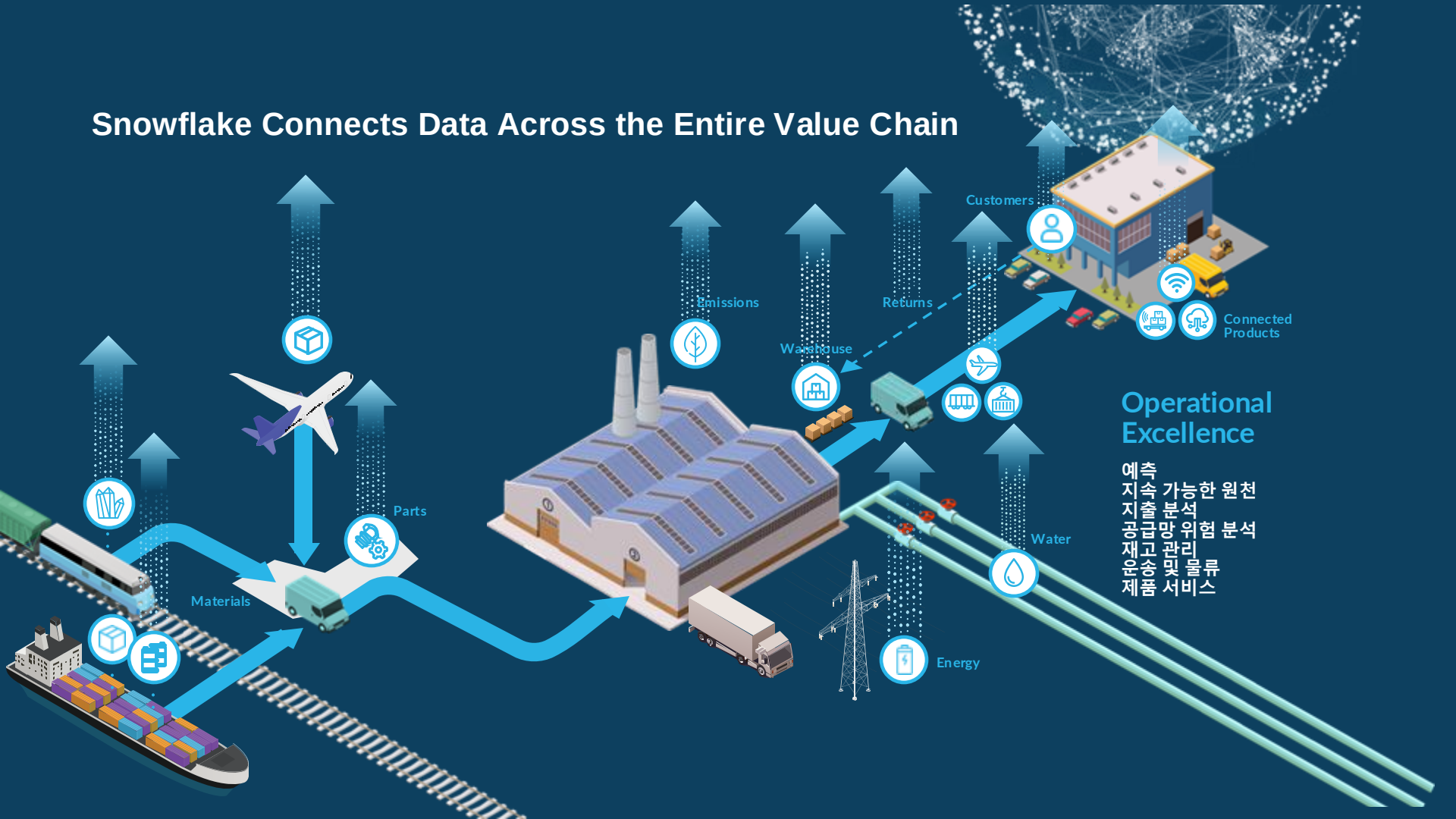
Haksoo Lee
Snowflake Korea

우리는 항상 AI 전략을 갖기 위해서는 데이터 전략이 필요하다고 말합니다.

Frank Slootman

Former **CEO Snowflake**

Snowflake Connects Data Across the Entire Value Chain




AI Data Cloud

Snowflake Evolution

Data Cloud Platform Innovation Journey

2012 — 2015 — 2019 — TODAY → **Platform for LLMs & Gen-AI**

CLOUD NATIVE ARCHITECTURE

Cloud Native한 아키텍처로 완전히 새롭게 만든 Data Warehouse로 유연하고 강력한 성능의 Cloud Data Warehouse

SINGLE DATA PALTFORM

단일 플랫폼으로 아키텍처를 간소화하고 데이터에 대한 사일로 제거

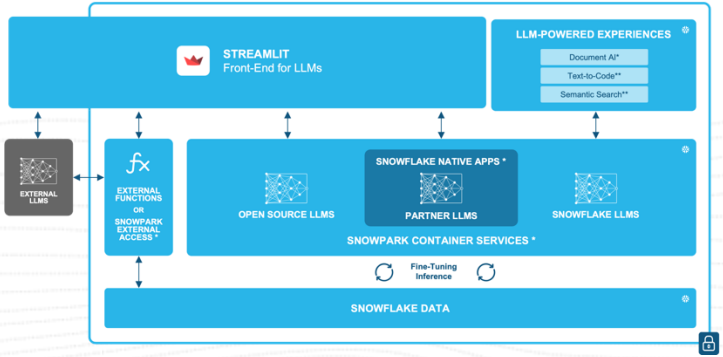
DATA SHARING & MONETIZATION

데이터 플랫폼 제공을 통하여 손 쉽게 데이터 뿐만 아니라 앱도 공유하고 이를 통한 수익화 모델 제공

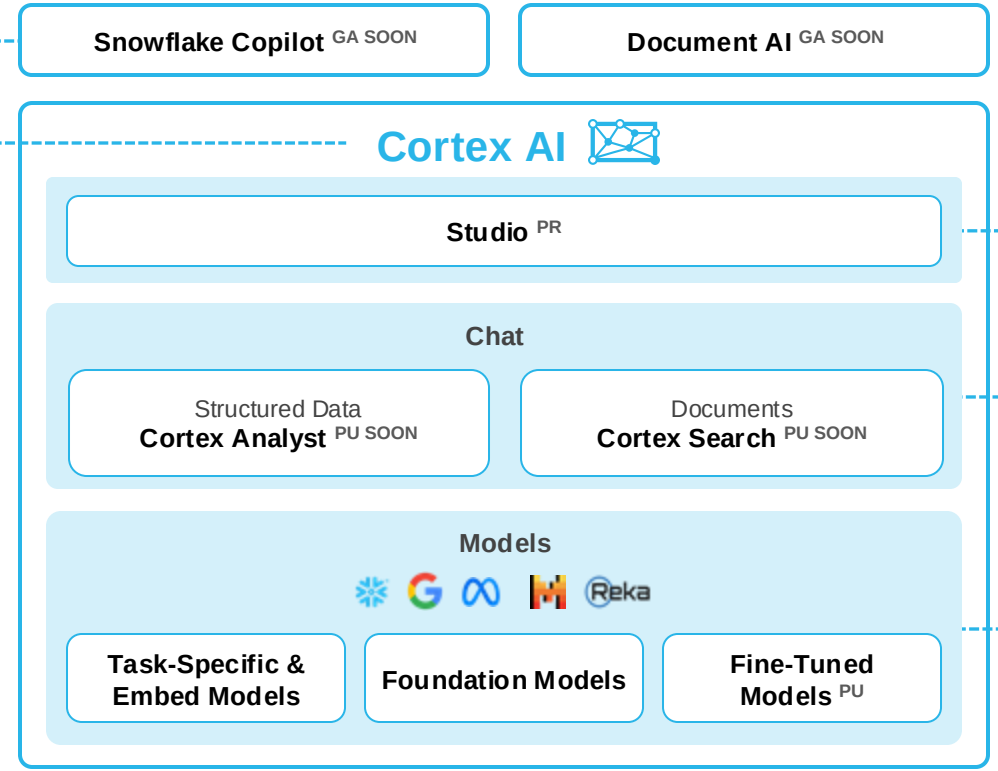
PROGRAMMABILITY

사용자 선호 프로그램 언어를 지원하여 거버넌스를 유지하면서 AI/ML 개발

SQL 및 Python 등의 언어를 알지 못하여도 누구든 자연어를 활용하여 질의를 하고 그에 맞는 결과 데이터를 쉽게 획득



Snowflake Gen AI Architecture

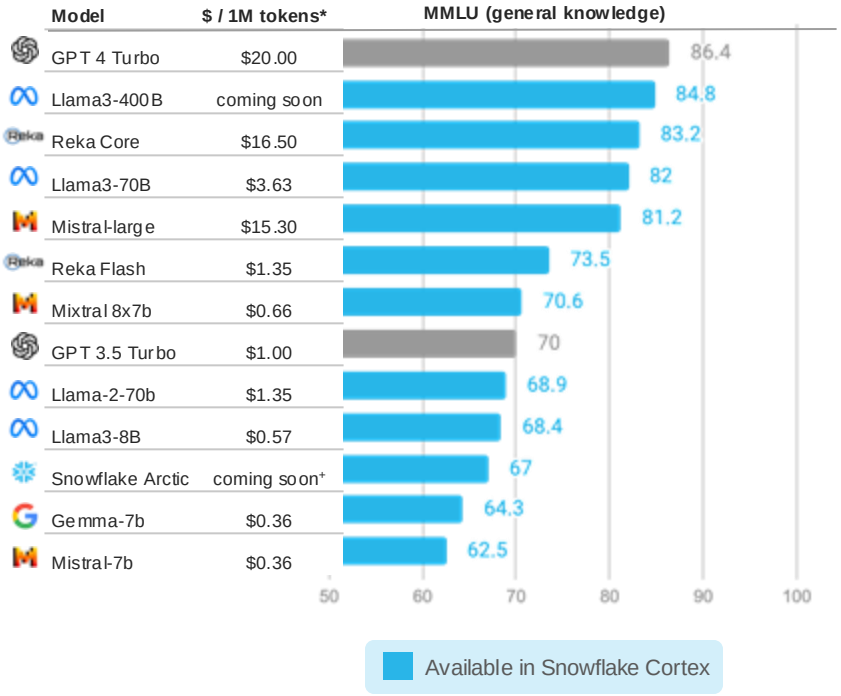


- **Snowflake Copilot**
  Snowflake Cortex 내에서 안전하게 실행되는 완전 관리형 서비스로 자연어 기반 데이터 분석 및 SQL 작성을 간편하게 도와주는 도구
- **Document AI**
  Snowflake에서 개발한 인공지능 모델(Arctic-TILT)로 다양한 문서 형식으로부터 정보를 추출할 수 있는 서비스 (계약서, 인보이스 등)
- **Cortex AI**
  Snowflake의 다양한 AI/ML 기능들의 집합으로 다양한 모델을 활용하여 정형 혹은 비정형 데이터를 이해하고 데이터 분석, 검색 및 예측 등의 AI 서비스를 손쉽게 구현할 수 있도록 서비스를 제공하는 AI 프레임워크

- **Studio**
  개발자가 아닌 일반 업무 담당자들도 AI 서비스를 코딩 없이 구축할 수 있도록 도와주는 Wizard
  - Fine-tune, Search(RAG)
  - 제공된 ML 기반의 예측, 이상탐지, 분류 등 분석 지원
- **Chat**
  정형/비정형 데이터에 대해 대화형 AI 서비스를 구축할 수 있도록 다양한 API 및 기능 제공
- **Cortex Analyst**
  구조화된 데이터를 업무 특성을 고려하여 자연어로 SQL 자동 작성 및 분석 가능 (Semantic Model 정의)
- **Cortex Search**
  Snowflake의 완전 관리형 RAG 기반 검색 서비스로, 벡터 및 키워드 검색을 결합한 고성능의 검색 엔진
- **Models**
  주요 모델 및 특정 목적에 따라 Fine-tuning된 모델 등 사전 검증된 다양한 모델 지원 (업무 특성 고려하여 유연하게 선택 가능)

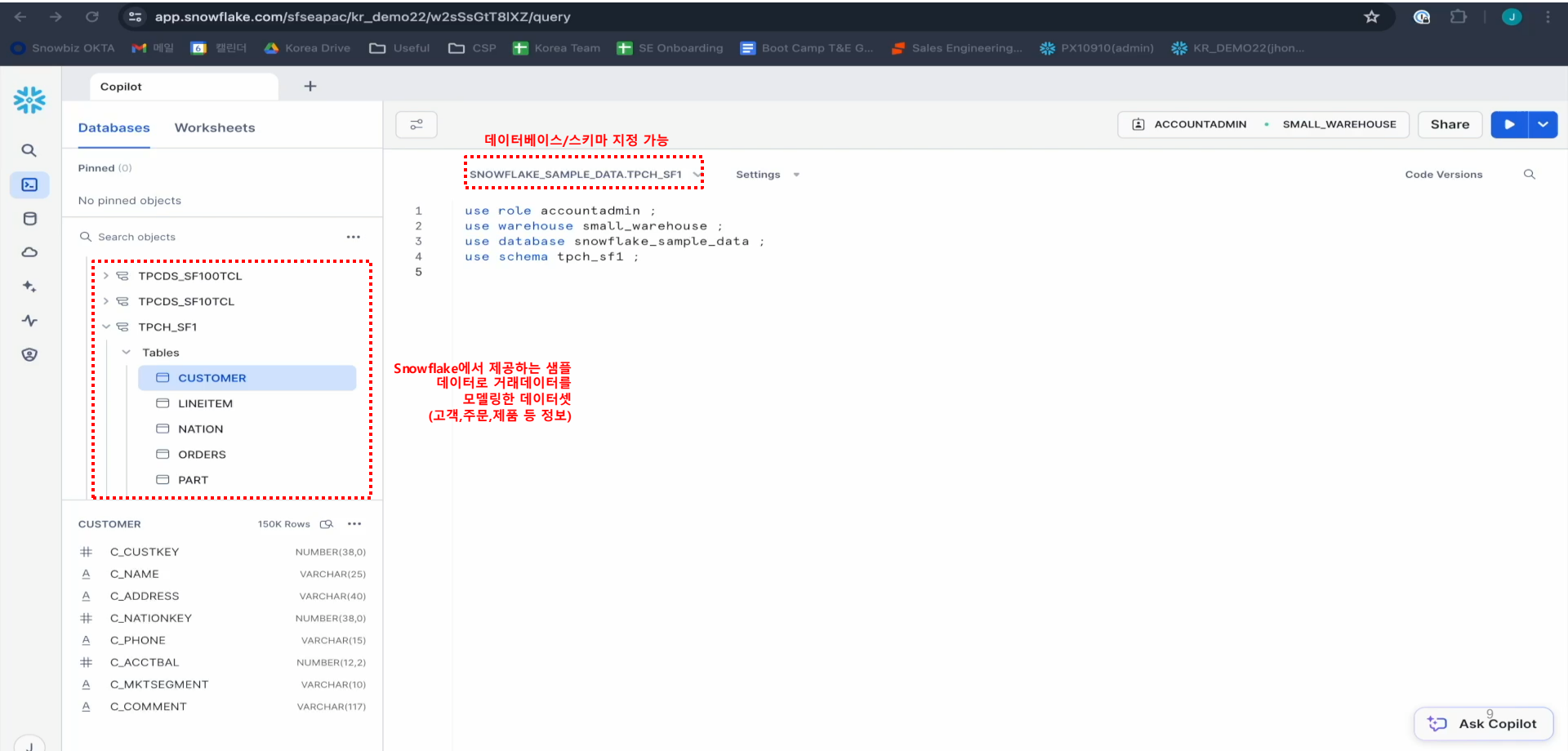
PR: Private Preview | PU: Public Preview | GA: General Availability

Foundation Models in Cortex AI



- 특정 작업과 목표를 위해 비용 및 성능, 정확성 등에 가장 최적화된 최신 모델들을 선택할 수 있는 유연성 제공
- Snowflake 내에서 서버리스 SQL 혹은 Python Function을 사용하여 배치 작업 실행 지원
- REST API를 통해 다양한 어플리케이션을 구축할 수 있는 개발 편의성 및 연동 지원
- GPU 등 AI 서비스 구축을 위한 인프라 관리 자동화
- 모든 데이터는 이동없이 Snowflake 내에서 처리함으로 안전한 데이터 운영 관리 가능
  (3rd Party LLM Provider와 공유되지 않음)
- RBAC 정책을 통한 역할에 따른 데이터 접근 제어 지원

Snowflake Copilot Demo



Cortex COMPLETE Function (1/2)

What Is It

서버리스 SQL / Python function 을 통해 모든 빌트인 모델로부터 효율적으로 추론을 실행

Why Use It

Snowflake 계정으로부터 데이터 이동 없이, 인프라를 관리할 필요 없이 Mistral AI, Llama, Google 의 LLMs 을 사용

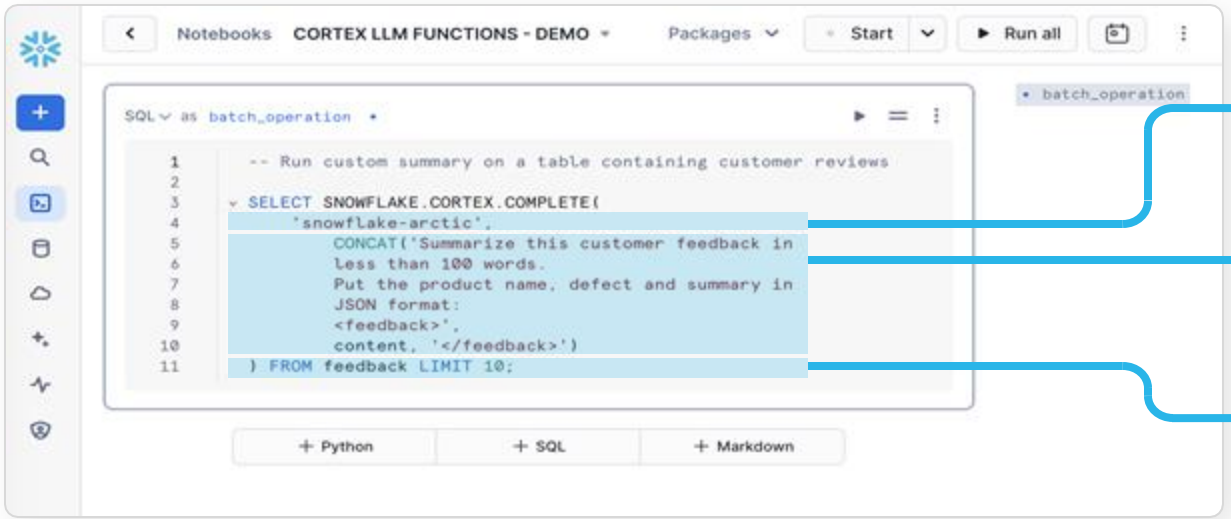
How To Use It

Worksheets/Notebooks 혹은 Streamlit 앱에서 SQL / Python 함수를 사용

Category	Model
Large	Mistral-Large
	Reka Core
Medium	Reka Flash
	Mixtral-8x7b
	Llama 70B
	Arctic
Small	Mistral-7B
	Gemma-7B
	Llama 3 8B

Cortex COMPLETE Function (2/2)

SQL을 통해 LLM 모델에 간편하게 질의하고 답변을 받을 수 있는 서버리스(Serverless) 함수를 제공합니다.



다양한 텍스트 처리 활용 사례:

- 사용자 정의 요약 작업 ex) text to JSON
- 카테고리별 분류 및 감정 분석 작업
- 이메일 및 컨텐츠 생성 작업

Available interfaces:

- SQL Statement
- Python Library
- REST API (PrPr)

Cortex FINETUNE Function

What Is It

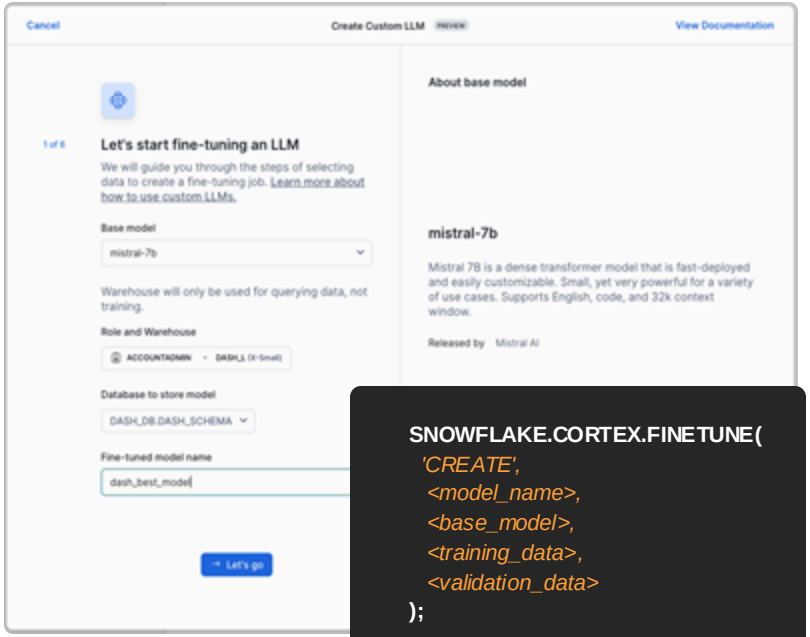
관리형 파인 튜닝 함수 : Cortex AI에서 사용할 수 있는 Mistral 및 Llama 3 파운데이션 모델을 대상으로 파인 튜닝 할 수 있는 함수 제공

Why Use It

- 파운데이션 LLM을 고객이 보유한 데이터를 활용해 파인 튜닝하여 모델 성능 및 정확도를 향상
- Gen AI Use Case에 대한 고품질 서비스 제공
- Snowflake 모델 레지스트리에 파인 튜닝한 모델을 등록하여 Custom LLM에 대한 액세스 및 거버넌스 관리

How To Use It

- Cortex Finetune()나 No-Code 인터페이스인 Studio를 통해 파인 튜닝을 수행할 수 있음.
- 파인 튜닝 된 LLM은 서버리스기반의 Cortex LLM Complete 함수에서 바로 사용 가능




```
SNOWFLAKE.CORTEX.FINETUNE(
  'CREATE',
    <model_name>,
    <base_model>,
    <training_data>,
    <validation_data>
);
```

Cortex Other Functions

특정 작업에 대해 다양한 함수들이 제공되며, 비용 효율적이고 우수한 성능을 보장합니다. (프롬프트 엔지니어링 불필요)

SUMMARIZE

주어진 텍스트로부터
커스터마이징이 필요 없는
빠른 문장 요약

SENTIMENT

고객 분석을 위해
고객 감정을 쉽게 감지
(긍정, 중립, 부정의 감정
지수로 응답)

EXTRACT_ANSWER

반정형/비정형 데이터에서
사용자 질의에 대한
응답을 추출
(신뢰도 점수 포함)

TRANSLATE

다양한 언어로
빠른 번역을 지원
(지원 가능한 언어는
지속적으로 개선되고 있음)

CLASSIFY_TEXT

자유형식의 문장을
사용자가 정의한
카테고리로 자동 분류

PARSE_DOCUMENT

PrPr

이미지로부터 텍스트
혹은 레이아웃(table)을
추출 **(OCR + ML)**

EMBED_TEXT_768 /EMBED_TEXT_1024

비정형 텍스트를 벡터
임베딩 처리하는 함수
(768, 1024 Dimension 지원)

COUNT_TOKENS

LLM Model 및 Function
사용 시 텍스트에 대한
토큰 수 확인

Document AI

What Is It

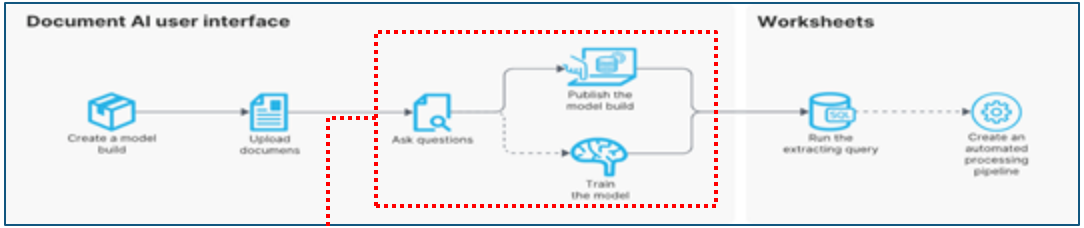
PDF 및 기타 비정형 문서에서 텍스트, 표 값, 수기 내용을 효율적으로 추출하기 위해 Arctic-TILT를 사용하는 완전 관리형 워크플로우

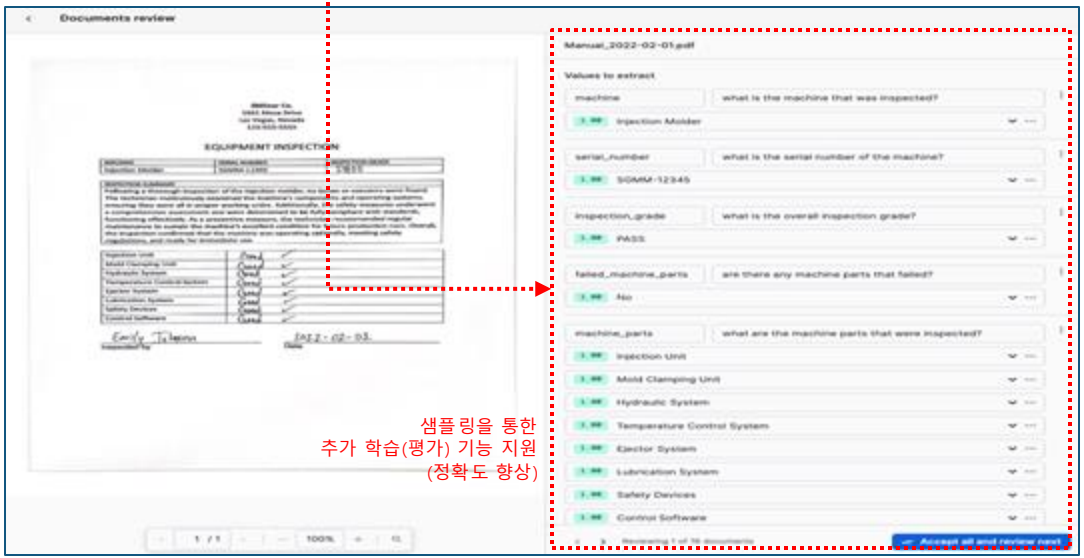
Why Use It

문서 처리에서 더 높은 효율성, 매뉴얼 작업 비용 감소 및 휴먼 에러 감소를 위해 업계 선도적인 LLM을 사용 - IDP

How To Use It

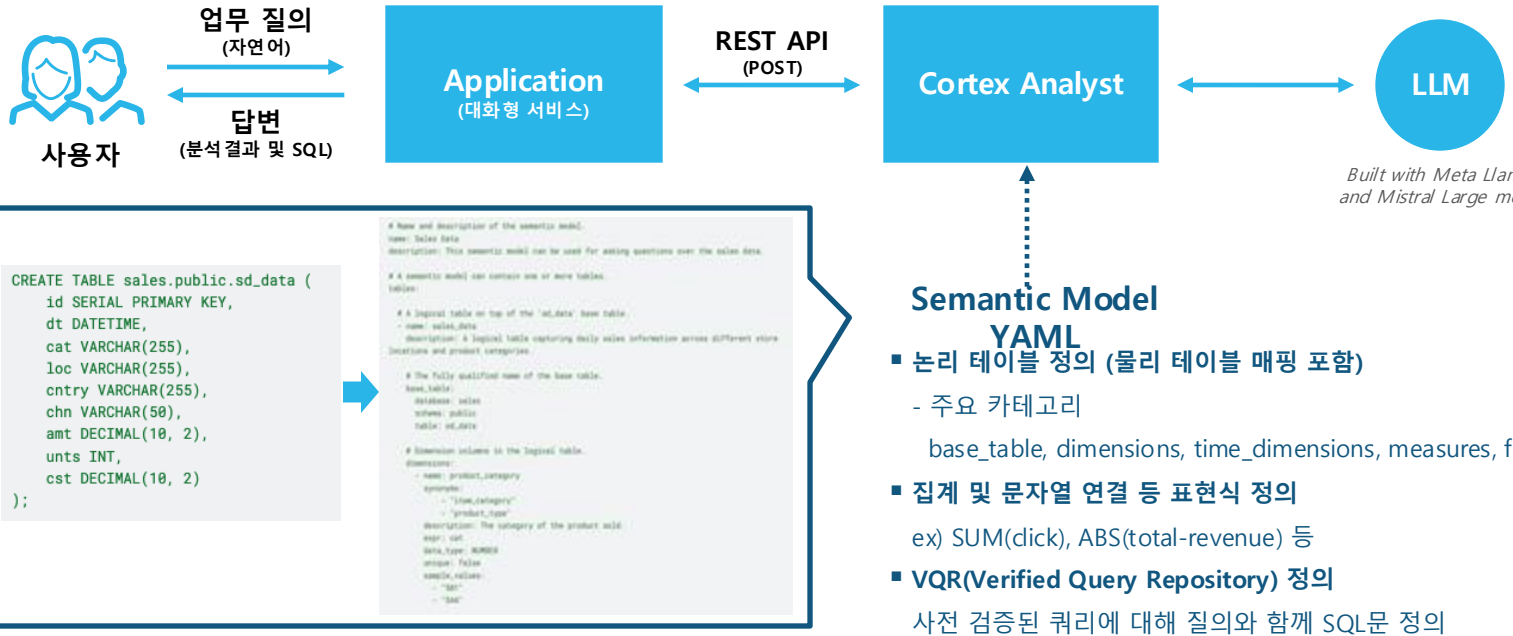
- **모델 준비 (비지니스 사용자)** : UI를 통한 자연어 기반 모델 생성(학습(Sampling)/ 평가) 및 게시 가능 - 코드 필요 없음
- **정보 추출 (데이터 팀)** : 준비된 모델 및 Stream, Task 등을 사용하여 파이프라인 자동화 구성





Cortex Analyst

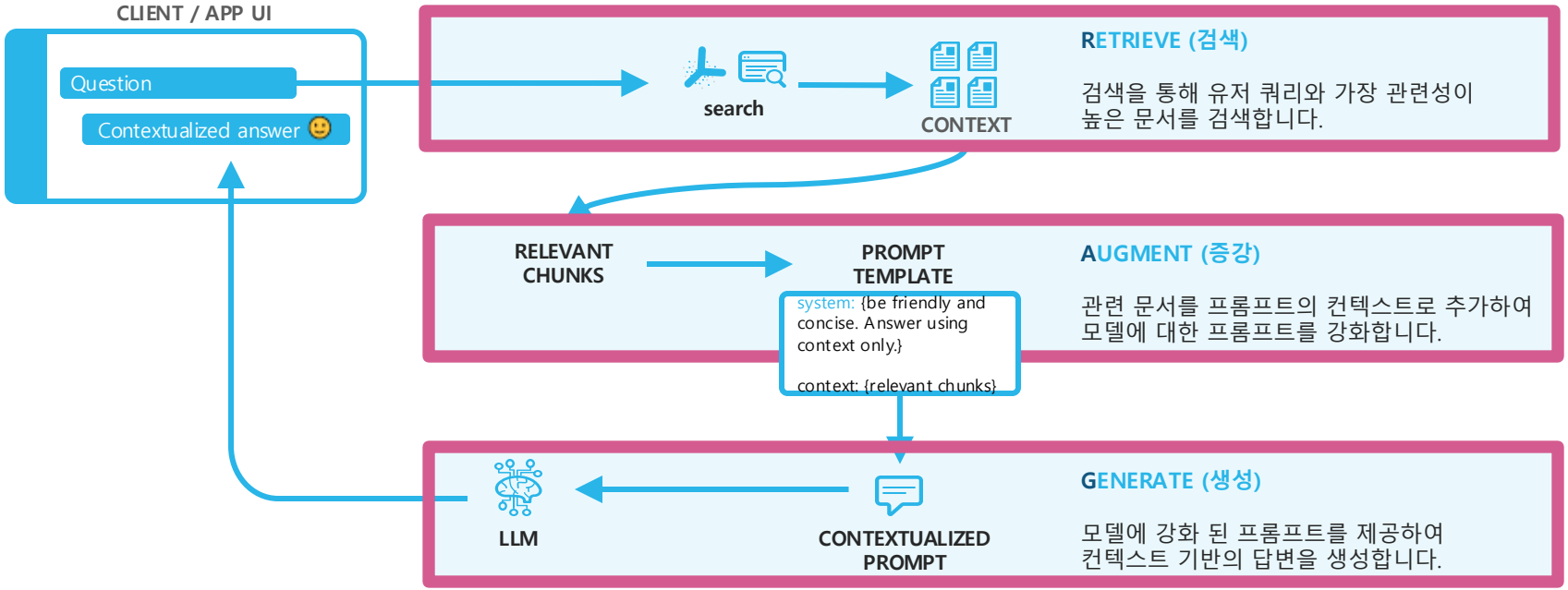
자연어를 통해 데이터 분석이 가능한 완전 관리형 LLM 기반 서비스로 특정 도메인에 대한 Semantic Model을 정의함으로 질의에 대한 답변의 정확도를 보다 높여줍니다. (유연한 REST API 접근 방식)



- **논리 테이블 정의 (물리 테이블 매핑 포함)**
  - 주요 카테고리
    base_table, dimensions, time_dimensions, measures, filters
- **집계 및 문자열 연결 등 표현식 정의**
  ex) SUM(click), ABS(total-revenue) 등
- **VQR(Verified Query Repository) 정의**
  사전 검증된 쿼리에 대해 질의와 함께 SQL문 정의
- **Semantic Model Generator (오픈소스)를 통한 자동 생성 가능**
  초기 YAML 파일 생성

Retrieval Augmented Generation (RAG)

- *기존 LLM 모델에 대해 정확성이나 신뢰성을 더욱 강화하기 위해 고안된 보완 기술*
- *특정 도메인에 대한 대량의 데이터를 관리하기 위한 Chunk, Embed, Vector 처리 등의 아키텍처 필요 (자동화)*



Cortex Search Overview

- ***완전 관리형 인덱싱 및 검색 (Fully managed indexing and retrieval) 서비스***

  - 임베딩, 인프라 관리, parameter 튜닝, 지속적인 인덱스 refresh 가 필요 없음

- ***대량의 데이터로부터 고성능, 고품질의 검색 성능을 제공하기 위해 RAG 기반의 하이브리드 검색 엔진 사용***

  ① 벡터 기반(의미적 유사도) 검색(Arctic-embed or multilingual model 사용) + 키워드 기반(어휘적 유사도) 검색(Neeva의 최첨단 웹 검색 기술 사용)

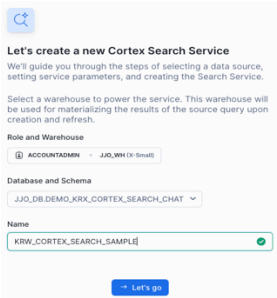
  ② Reranking → Score Fusion 단계를 거쳐 질의에 대한 가장 관련성이 높은 컨텍스트 세트를 출력

- ***벡터 데이터 구성(Embed Processing)을 위한 복잡한 절차를 간소화하여 생성 및 관리할 수 있는 기능 제공***

  - Wizard 및 SQL을 통한 서비스 손쉽게 정의

  - 유연한 서비스 연동을 위한 다양한 호출 방법 제공

Snowsight(Wizard)



SQL Statement

```
CREATE CORTEX SEARCH SERVICE KRW_CORTEX_SEARCH_SAMPLE
  ON DESCRIPTION
  ATTRIBUTES CITY, REGION
  WAREHOUSE = KRX_WH
  TARGET_LAG = '1 day'
  AS (
    SELECT
        CITY,
        REGION,
        DESCRIPTION
    FROM PREQIN_INVESTOR_INFO
);
```

Python Library

```
# fetch service
my_service = (root
  .databases["<service_database>"]
  .schemas["<service_schema>"]
  .cortex_search_services["<service_name>"]
)

# query service
resp = my_service.search(
  query="<query>",
  columns=["<col1>", "<col2>"],
  filter={"@eq": {"<column>": "<value>"} },
  limit=5
)
print(resp.to_json())
```

REST API

```
curl --location https://<ACCOUNT_URL>/api/v2/databases/<DB_NAME
--header 'X-Snowflake-Authorization-Token-Type: KEYPAIR_JWT' \
--header 'Content-Type: application/json' \
--header 'Accept: application/json' \
--header "Authorization: Bearer $CORTEX_SEARCH_JWT" \
--data '{
  "query": "<search_query>",
  "columns": ["col1", "col2"],
  "filter": <filter>
  "limit": <limit>
}'
```

SQL Function (PrPr)

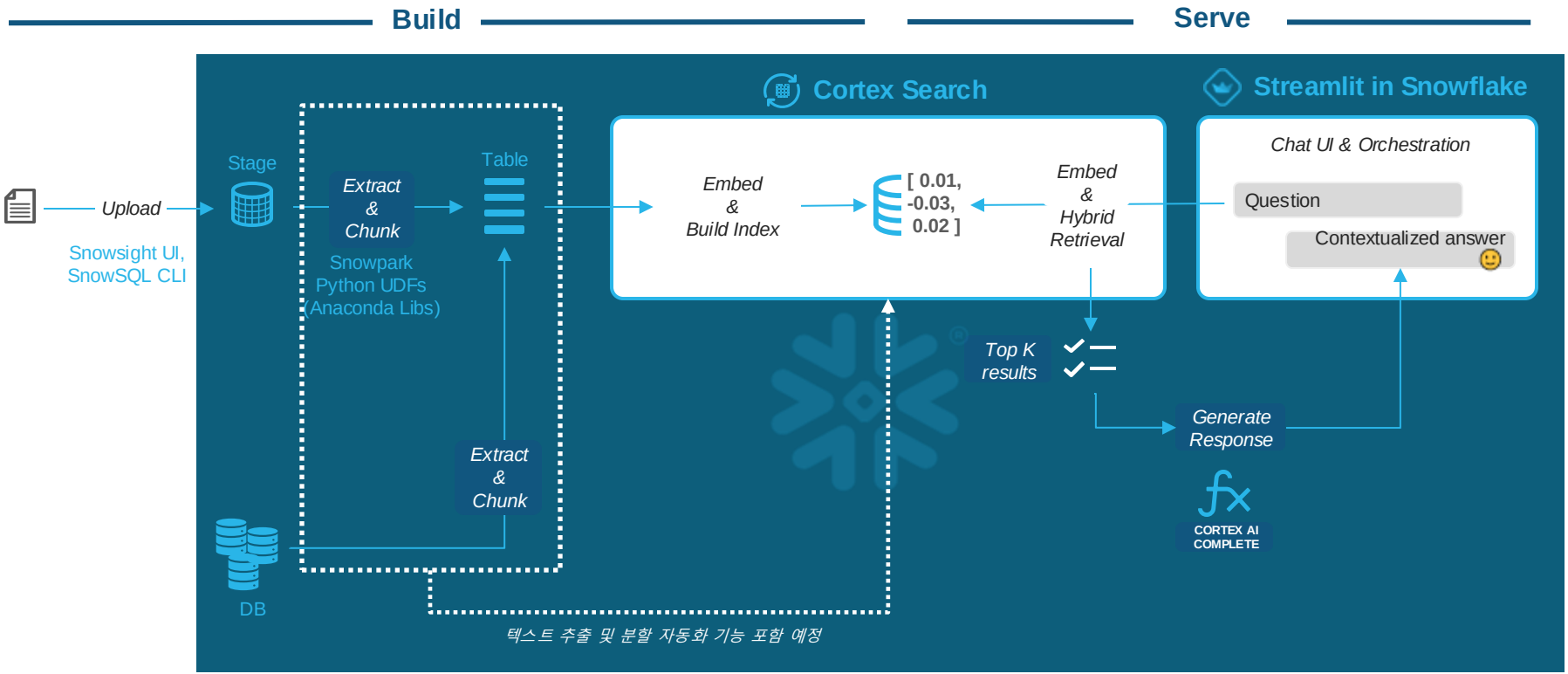
```
SELECT PARSE_JSON(
  SNOWFLAKE.CORTEX.SEARCH_PREVIEW(
      'my_search_service',
      '{
         "query": "preview query",
         "columns":[
            "col1",
            "col2"
         ],
         "filter": {"@eq": {"col1": "filter value"} },
         "limit":10
      }'
  )
)['results'] as results;
```

- ***주요 사용 사례***

  - 엔터프라이즈 검색 서비스 : 고성능/고품질의 검색 기능을 활용한 검색 서비스로 활용

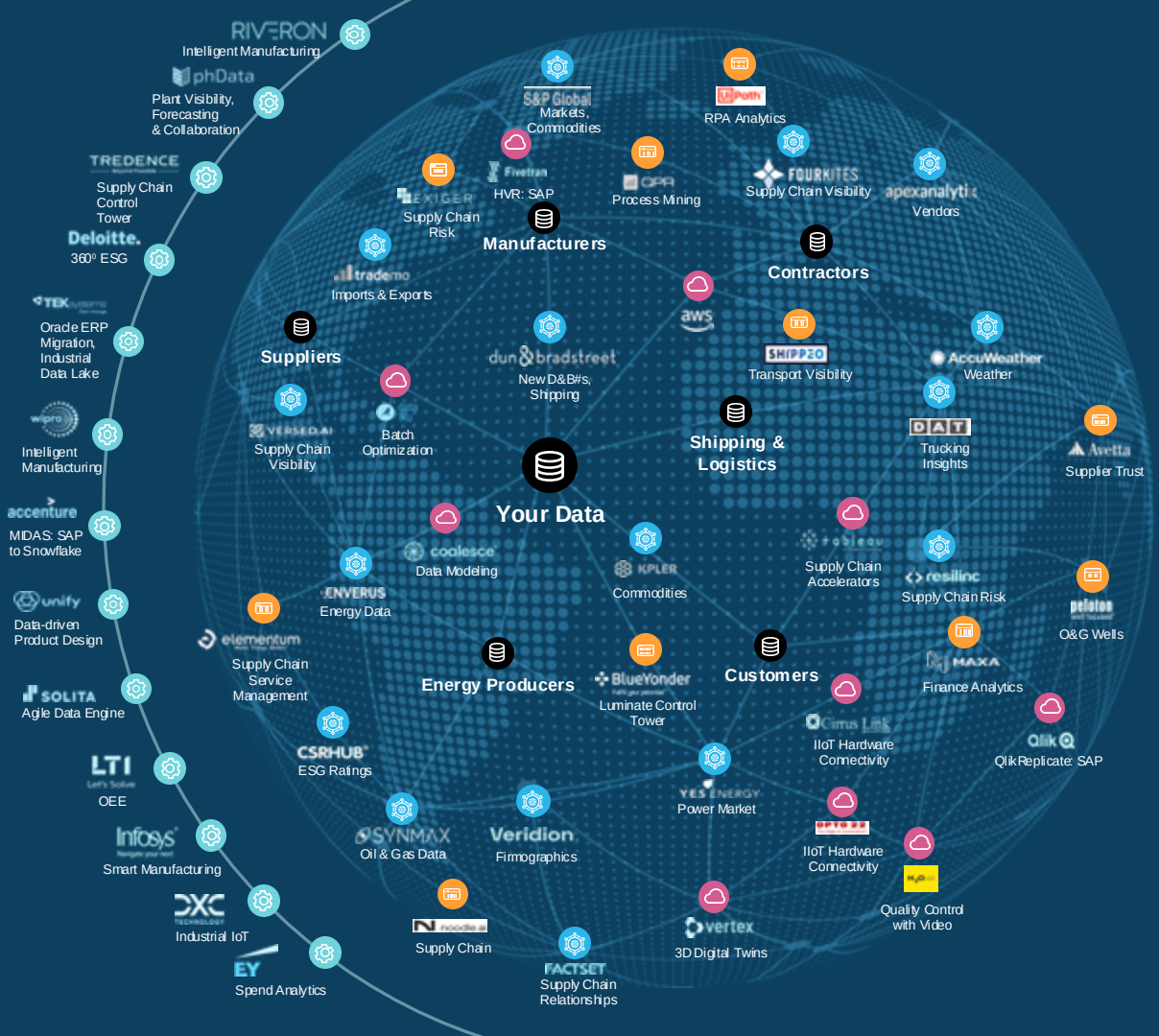
  - 특정 도메인에 대한 Chatbot 서비스 적용 : 맞춤별, 상황별 응답을 위한 의미론적 검색을 활용하여, 자연어 기반의 대화형 어플리케이션에 검색 엔진으로 활용

Cortex Search 아키텍처(RAG 엔진)



제조산업 데이터 클라우드

- 산업데이터 마켓플레이스
- Powered by Snowflake
- 기술 파트너
- SI 파트너



Customer Use Case

About ABB

ABB는 보다 생산적이고 지속 가능한 미래를 달성하기 위해 사회와 산업의 변화에 활력을 불어넣는 선도적인 글로벌 기술 회사입니다.
ABB는 소프트웨어를 전기화, 로봇 공학, 자동화 및 모션 포트폴리오에 연결함으로써 기술의 경계를 넓혀 성능을 새로운 수준으로 끌어올립니다.
ABB의 전기화 사업은 변전소에서 소켓에 이르기까지 광범위한 제품, 디지털 솔루션 및 서비스 포트폴리오를 제공하여 안전하고 스마트하며 지속 가능한 전기화를 가능하게 합니다. 제품에는 EV 인프라, 태양광 인버터, 모듈식 변전소, 배전 자동화, 전력 보호, 배선 액세서리, 개폐 장치, 인클로저, 케이블링, 감지 및 제어를 포함한 저전압 및 중전압을 위한 디지털 및 커넥티드 혁신이 포함됩니다.

Industry:

Electrical Manufacturing

Countries:

100+

Employees:

105,000

Years of operation:

130

Orders:

~$32 bn

Customer Use Case

Challenges

산재된 환경

ABB에는 100 + ERP 시스템과 수천 개의 사용자 정의 응용 프로그램에 대한 일종의 데이터 분석을 제공하는 35 개 이상의 시스템이있었습니다.

- IT 환경에서 초점 또는 우선 순위에 대한 조정이 거의 또는 전혀 없음
- IT에는 로컬 및 / 또는 글로벌 수준의 소유자가있었습니다.
- 조명을 유지하기 위해 ~ 250 명의 팀이 필요했습니다.
- 실제 분석 솔루션에 소요되는 노력이 거의 없습니다.

시간 소모적인 RCA

산재된 IT환경과 수많은 개별 조직들은 원인분석과정에 필요한 데이터 이슈를 해결하기 위해 광범위한 노력이 필요함.

RCA 프로세스는 각 팀이 경쟁하는 우선 순위와 개별 릴리스 일정을 가지고 있었기 때문에 조정하기가 어려웠고, 이로 인해 MTTR이 연장되었습니다.

조직 변화

ABB는 국가별 운영모델에서 전세개를 4개 비즈니스 권역으로 통합하는 작업을 수행했습니다.

이 과정에서 데이터가 지리적으로 분산되어 있기 때문에 각 비즈니스 권역에서 싱글뷰를 제공하는데 상당한 문제가 발생했습니다.

Customer Use Case

Snowflake Use Case

생산계획 최적화

전기화 공장 및 지역 배송 센터의 생산 관리자 및 분석가는 수동 보고서 작성 및 RCA 수행과 같은 소모적인 업무에 시간의 20-30%를 소비합니다.

보고서는 종종 신뢰할 수 없고, 오류가 발생하기 쉬우며, 특정 위치 밖에서는 사용할 수 없었습니다. 남은 시간의 대부분은 취소를 관리하고, 주문을 처리하고, 쓸모없는 재고를 관리하는 업무를 수행하는데 소비했습니다.

Material Stabilization Target:

95%

재고와 생산관리에 대한 통찰력

계획 수립 담당자 및 관리자들이 취소, 긴급 주문, 예외 및 재고 조정을 줄이는 조치를 트리거할 수 있도록 지원

생산성 향상

매뉴얼 보고절차를 없애, 해당 팀과 분석가들이 의사결정 최적화에 더 많은 시간을 할애

향상된 투명성 및 속도

Provide true visibility across all levels of the organization into material forecasts, shortages and excess inventory risks. Decrease time to action

조직의 모든 구성원에게 자재 예측, 부족 및 초과 재고위험에 대한 투명한 가시성을 제공함으로써 MTTA를 단축.

THANK YOU