▣ **강의명:** CSCI E-103: 재현 가능한 머신러닝

▣ **주차:** Lecture 14

▣ **교수명:** Anindita Mahapatra
Eric Gieseke

▣ **목적:** Lecture 14의 핵심 개념 학습

▣ **강의명:** Big Data Analysis & Engineering

▣ **주차:** Day 14 - Databricks Roadmap & Review

▣ **교수명:** Prof. & Teaching Staff

▣ **목적:** Databricks의 최신 기능(Lakebase, AI/BI, Agent) 이해 및 Quiz 2, Assignment 3의 주요 개념(Spark 최적화, Streaming) 심층 리뷰

Contents

1 Databricks Data Intelligence Platform (Roadmap)

이번 강의에서는 Databricks 로드맵에 포함된 최신 기능들을 다룹니다. 이 기능들은 주로 **Data Intelligence Capabilities**를 강화하여 플랫폼을 더 스마트하게 만들고 데이터 담당자의 업무를 줄이는 데 초점을 맞춥니다.

1.1 주요 구성 요소 요약

Category	Feature	Description
Data Warehousing	Lakebase (Postgres)	Lakehouse 환경 내에서 OLTP 트랜잭션 처리를 지원하는 관리형 Postgres.
AI & Agents	Agent Bricks	AutoML처럼 데이터셋만 지정하면 에이전트를 자동 생성해주는 프레임워크.
	Genie	정형화된 "What" 질문에 대한 답변을 제공하는 BI 도구.
	Research Mode	가설 기반의 "Why" 질문에 대해 Chain of Thought를 통해 심층 답변 제공.
Governance	ABAC	속성 기반 접근 제어(Attribute-Based Access Control). PII 태그 등에 따라 자동 권한 부여.
Apps	Lakehouse Apps	데이터가 있는 곳에서 바로 실행되는 Python(Streamlit 등) 기반 앱. 보안 및 확장성 보장.

Table 1: *Databricks* 주요 신규 기능 요약

1.2 Lakebase (Managed Postgres)

기존의 Databricks는 OLAP(분석용)에 최적화되어 있었으나, **Lakebase**의 도입으로 OLTP(트랜잭션용) 워크로드까지 커버하게 되었습니다.

- **구조적 특징**: Compute와 Storage가 분리되어 있습니다.
- **장점**:
  - 사용량이 늘어나면 Compute가 확장(Scale-up)되고, 사용하지 않으면 0으로 축소(Scale-to-Zero)되어 비용 효율적입니다.
  - Production과 Development 브랜치를 기본 제공하여, 운영 환경에 영향을 주지 않고 테스트가 가능합니다.
- **통합**: Unity Catalog와 연동되며, Lakehouse와 Lakebase 간 양방향 데이터 동기화(Reverse Sync)가 가능합니다.

1.3 AI/BI: Genie & Research Mode

- **Genie**: 정형 데이터에 대한 사실적 질문(Fact-based questions)을 SQL로 변환하여 답변합니다.
- **Research Mode**: "왜(Why)"에 대한 질문을 처리합니다. 단순 검색이 아니라 가설을 세우고 여러 경로를 탐색(Reasoning)하여 답변을 도출합니다. (예: "관세 인상이 포트폴리오에 미칠 영향은?")

1.4 Apps & Model Serving

- **Apps**: Streamlit, Flask 등으로 만든 앱을 데이터가 저장된 플랫폼 위에서 직접 구동합니다. 데이터 이동 없이 보안이 유지된 상태로 대시보드 이상의 상호작용(Write-back 등)이 가능합니다.
- **Model Serving**: 최신 LLM(Llama, Gemini, Claude 등)을 즉시 사용할 수 있도록 제공하며, MCP(Model Context Protocol) 서버를 통해 에이전트가 외부 도구나 데이터에 접근하도록 지원합니다.

2 Quiz 2 Review: Spark & Delta Lake Concepts

퀴즈에서 많이 혼동했던 Spark와 Delta Lake의 핵심 개념을 정리합니다.

2.1 Spark Optimization & Troubleshooting

1. **Spark 최적화 4대 요소**: Data Skew, Shuffle, Spill, Small Files.
2. **Broadcast Variables**:
   - **Immutable (불변)**: 생성 후 변경할 수 없습니다.
   - **Local to Workers**: 클러스터 간 공유되는 것이 아니라, 각 워커 노드에 복사본이 저장됩니다.
3. **Repartition vs Coalesce**:
   - 파티션을 **늘릴 때** (예: 12 → 24): 'repartition()' 사용 (Shuffle 발생).
   - 파티션을 **줄일 때**: 'coalesce()' 사용 (Shuffle 최소화).
4. **Action 함수 주의사항**:
   - 'collect()': 모든 데이터를 드라이버로 가져오므로 OOM(Out of Memory) 위험이 큼. 디버깅용으로만 사용 권장.
   - 'take(n)': 필요한 n개 행만 가져오므로 대용량 데이터에서 안전함.

2.2 Delta Lake Internals

- **Transaction Log**: '$_delta_log$' .
- **구조**: 각 트랜잭션은 **개별 JSON 파일**로 기록됩니다. (하나의 JSON이 아님). 10번째 트랜잭션마다 Checkpoint(Parquet) 파일이 생성됩니다.
- **Time Travel**: 별도의 복사본을 만드는 것이 아니라, 트랜잭션 로그를 통해 과거 시점의 데이터 상태를 조회합니다.

2.3 Data Frame Operations

- **Explode**: DataFrame의 메서드가 아니라 'select' 문 내에서 사용해야 합니다. (예: 'df.select(explode(col))')
- **Drop Duplicates**: 'df.dropDuplicates(['col1', 'col2'])' 형태로 리스트를 전달해야 합니다.

3 Assignment 3 Review: Streaming & Architecture

3.1 Kafka vs. Kinesis

Feature	AWS Kinesis	Apache Kafka
관리 주체	AWS 완전 관리형 (Managed)	Confluent (상용) 또는 Open Source (직접 관리)
확장 단위	**Shard** (샤드 수에 따라 비용/성능 결정)	**Partition** (파티션 단위로 병렬 처리)
데이터 분배	Partition Key 사용	Round-robin 또는 Key-Value 기반 Partitioning

Table 2: *Kinesis*와 *Kafka* 비교

3.2 Slowly Changing Dimensions (SCD)

- **Type 1**: 과거 데이터를 덮어씀(Overwrite). 이력 관리 안 됨.
- **Type 2**: 새로운 행을 추가하고 유효 기간(Start/End Date)이나 플래그(Current)를 두어 이력을 관리함.
- **Delta Merge**: 'MERGE INTO' 구문을 사용하여 변경된 컬럼(예: City)만 감지하고 업데이트/삽입 로직을 구현합니다.

3.3 Spark Streaming Logic

- **Trigger 옵션**:
  - 'trigger(once=True)': **Deprecated**.
  - 'trigger(availableNow=True)': 권장됨. 데이터를 마이크로 배치로 나누어 처리하므로 클러스터 과부하를 방지합니다.
- **Checkpointing**: 스트리밍의 핵심. 시스템 중단 후 재시작 시 중단된 지점부터 정확히 다시 처리(Exactly-once)할 수 있게 해줍니다.
- **Image Processing**:
  - Spark Serverless 환경에서는 'display()'로 이미지를 직접 렌더링하지 못할 수 있습니다.
  - 'pillow' 라이브러리 등을 사용하여 바이너리 데이터를 이미지로 변환하거나 처리(Invert 등)하는 UDF를 작성해야 합니다.

4 Next Steps & Action Items

- **Assignment 4**: 다음 수업 전까지 최종 과제(Assignment 4)를 확인하고 준비할 것.
- **Guest Lecture**: 다음 수업에는 업계 전문가 초청 강연이 예정되어 있음.