**AZ-실시간에 가까운 레이크하우스 데이터 처리

실시간에 가까운 레이크하우스 데이터 처리

Azure AI 검색

Azure 코스모스 DB

Azure 데이터 레이크

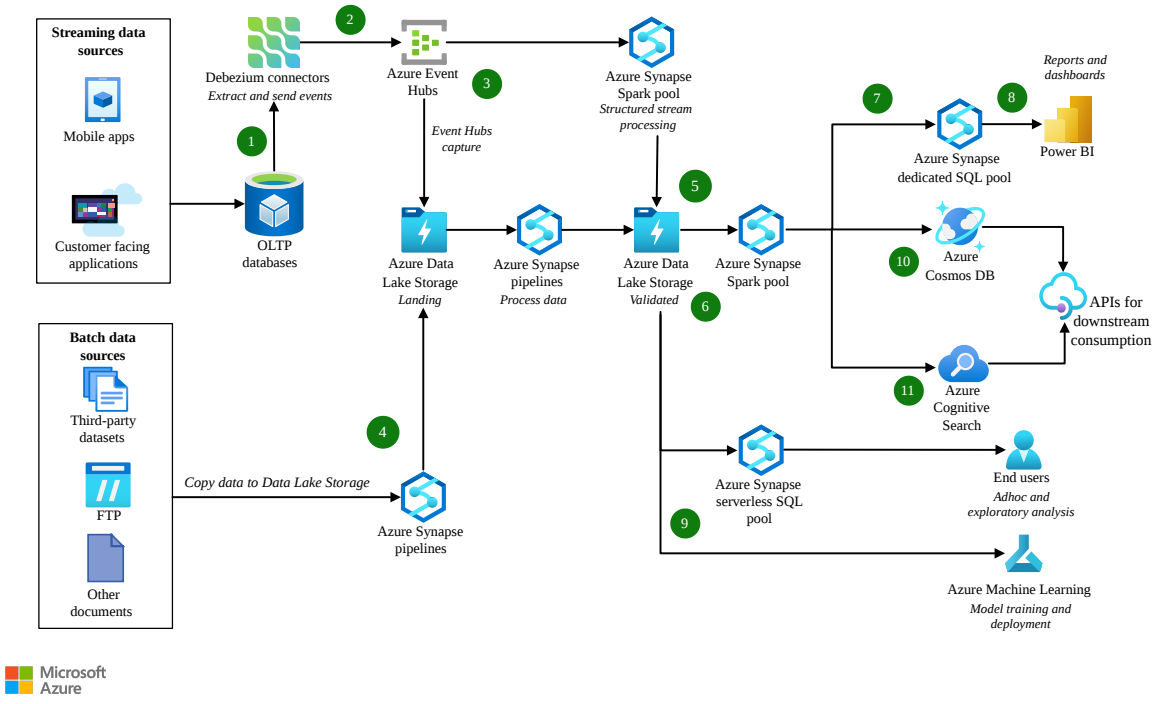
Azure 이벤트 허브

Azure Synapse 분석

데이터 중심 기업은 백엔드 및 분석 시스템을 고객 대면 애플리케이션과 거의 실시간으로 동기화해야 합니다. 거래, 업데이트 및 변경의 영향은 엔드투엔드 프로세스, 관련 애플리케이션 및 온라인 거래 처리(OLTP) 시스템을 통해 정확하게 반영되어야 합니다. OLTP 애플리케이션의 변경 사항이 데이터를 사용하는 다운스트림 시스템에 반영되는 데 허용되는 지연 시간은 몇 분에 불과할 수 있습니다.

이 문서에서는 레이크하우스 데이터를 동기화 상태로 유지하기 위한 거의 실시간 데이터 처리를 위한 엔드투엔드 솔루션을 설명합니다. 이 솔루션은 데이터 처리 및 분석을 위해 Azure Event Hubs, Azure Synapse Analytics, Azure Data Lake Storage를 사용합니다.

건축학



이 아키텍처의 Visio 파일을 다운로드하세요 .

데이터 흐름

1. 변경 데이터 캡처는 소스 시스템이 변경 사항을 수신하기 위한 전제 조건입니다. Debezium 커넥터는 다양한 소스 시스템에 연결하여 변경 사항이 발생하는 대로 이를 활용할 수 있습니다. 커넥터는 변경 사항을 캡처하고 다양한 관계형 데이터베이스 관리 시스템(RDBMS)에서 이벤트를 생성할 수 있습니다. Debezium 커넥터를 설치하려면 Kafka 연결 시스템이 필요합니다.
2. 커넥터는 변경 데이터를 추출하고 캡처된 이벤트를 Azure Event Hubs로 전송합니다. Event Hubs는 여러 소스에서 대량의 데이터를 수신할 수 있습니다.
3. Event Hubs는 데이터를 Azure Synapse Analytics Spark 풀로 직접 스트리밍하거나 원시 형식으로 Azure Data Lake Storage 랜딩 존으로 전송할 수 있습니다.
4. 다른 배치 데이터 소스는 Azure Synapse 파이프라인을 사용하여 데이터를 Data Lake Storage에 복사하고 처리할 수 있도록 할 수 있습니다. 엔드투엔드 ETL(추출, 변환 및 로드) 워크플로는 여러 단계를 연결하거나 단계 간에 종속성을 추가해야 할 수 있습니다. Azure Synapse 파이프라인은 전체 처리 프레임워크 내에서 워크플로 종속성을 조정할 수 있습니다.
5. Azure Synapse Spark 풀은 완벽하게 지원되는 Apache Spark 구조화된 스트리밍 API를 사용하여 Spark 스트리밍 프레임워크에서 데이터를 처리합니다. 데이터 처리 단계는 데이터 품질 검사와 고수준 비즈니스 규칙 검증을 통합합니다.

6. Data Lake Storage는 검증된 데이터를 오픈 델타 레이크 포맷으로 저장합니다. 델타 레이크는 원자성, 일관성, 격리성, 내구성(ACID) 의미론과 트랜잭션, 확장 가능한 메타데이터 처리, 기존 데이터 레이크에 대한 통합 스트리밍 및 일괄 데이터 처리를 제공합니다.

   쿼리 가속을 위해 인덱스를 사용하면 델타가 더욱 향상된 성능을 제공합니다. Data Lake Storage 검증 영역의 데이터는 더욱 진보된 분석 및 머신 러닝의 소스가 될 수도 있습니다.

7. Data Lake Storage 검증 영역의 데이터는 더 많은 규칙으로 변환되고 보강되어 최종 처리된 상태로 제공되며, 대규모 분석 쿼리를 실행하기 위해 전용 SQL 풀에 로드됩니다.

8. Power BI는 전용 SQL 풀을 통해 노출된 데이터를 사용하여 엔터프라이즈급 대시보드와 보고서를 구축합니다.

9. Data Lake Store 랜딩 존에서 캡처한 원시 데이터와 Delta 형식의 검증된 데이터를 다음에 사용할 수도 있습니다.

   - Azure Synapse SQL 서버리스 풀을 통한 추가적인 임시 및 탐색적 분석.
   - Azure Machine Learning을 통한 머신 러닝.

10. 일부 저지연 인터페이스의 경우, 데이터는 단일 자릿수 서버 지연에 대해 비정규화되어야 합니다. 이 사용 시나리오는 주로 API 응답을 위한 것입니다. 이 시나리오는 Azure Cosmos DB와 같은 NoSQL 데이터 저장소에서 문서를 쿼리하여 단일 자릿수 밀리초 응답을 얻습니다.

11. Azure Cosmos DB 파티셔닝 전략은 모든 쿼리 패턴에 적합하지 않을 수 있습니다. 그런 경우 API가 Azure Cognitive Search로 액세스해야 하는 데이터를 인덱싱하여 솔루션을 보강할 수 있습니다. Azure Cosmos DB와 Cognitive Search는 대기 시간이 짧은 쿼리 응답이 필요한 대부분의 시나리오를 충족할 수 있습니다.

구성 요소

이 솔루션은 다음과 같은 Azure 구성 요소를 사용합니다.

- Event Hubs 는 방대한 양의 데이터를 수집하도록 확장할 수 있는 관리형 분산 수집 서비스입니다. Event Hubs 구독자-게시자 메커니즘을 사용하면 다양한 애플리케이션이 Event Hubs의 토픽에 메시지를 보낼 수 있고, 다운스트림 소비자는 메시지에 연결하여 처리할 수 있습니다. Event Hubs 캡처 기능은 메시지가 도착하면 AVRO 형식으로 Data Lake Storage에 메시지를 쓸 수 있습니다. 이 기능을 통해 간편한 마이크로 배치 처리 및 장기 보관 시나리오가 가능합니다. Event Hubs는 또한 Kafka 호환 API를 제공하고 스키마 레지스트리를 지원합니다.

- Data Lake Storage는 모든 데이터를 원시 및 검증된 형식으로 저장하는 스토리지 하위 시스템을 형성합니다. Data Lake Storage는 대규모로 트랜잭션을 처리할 수 있으며, 다양한 파일 형식과 크기를 지원합니다. 계층적 네임스페이스는 데이터를 익숙한 폴더 구조로 구성하는 데 도움이 되며, UniX(POSIX) 권한을 위한 Portable Operating System Interface를 지원합니다. Azure Blob Filesystem(ABFS) 드라이버는 Hadoop 호환 API를 제공합니다.
- Azure Synapse Analytics 는 데이터 통합, 엔터프라이즈 데이터 웨어하우징, 빅데이터 분석을 하나로 결합한 무한 분석 서비스입니다. 이 솔루션은 Azure Synapse Analytics 에코시스템의 다음 기능을 사용합니다.
    - Azure Synapse Spark 풀은 오픈소스 Spark에 기본 제공 성능 향상을 추가하는 주문형 Spark 런타임을 제공합니다. 고객은 유연한 자동 크기 조정 설정을 구성하고, Apache Livy 엔드포인트를 통해 원격으로 작업을 제출하고, Synapse Studio 노트북 인터페이스를 사용하여 대화형 환경을 만들 수 있습니다.
    - Azure Synapse SQL 서버리스 풀은 익숙한 T-SQL 구문을 사용하여 레이크하우스 데이터를 쿼리하는 인터페이스를 제공합니다. 설정할 인프라가 없으며 Azure Synapse 작업 영역 배포는 자동으로 엔드포인트를 만듭니다. Azure Synapse SQL 서버리스 풀은 제자리에서 데이터의 기본 검색 및 탐색을 가능하게 하며 사용자 임시 쿼리 분석에 적합한 옵션입니다.
    - Azure Synapse 전용 SQL 풀은 열 저장소가 있는 관계형 테이블에 데이터를 저장합니다. 전용 SQL 풀은 확장형 아키텍처를 사용하여 여러 노드에 데이터 처리를 분산합니다. PolyBase 쿼리는 데이터를 SQL 풀 테이블로 가져옵니다. 테이블은 분석 및 보고를 위해 Power BI에 연결할 수 있습니다.
- Power BI는 보고서와 대시보드를 만들고 액세스하기 위한 시각적 인터페이스를 제공합니다. Power BI Desktop은 다양한 데이터 소스에 연결하고, 소스를 데이터 모델로 결합하고, 보고서나 대시보드를 빌드할 수 있습니다. Power BI를 사용하면 비즈니스 요구 사항에 따라 데이터를 변환하고, Power BI 서비스를 통해 다른 사람과 비주얼과 보고서를 공유할 수 있습니다.
- Azure Cosmos DB는 MongoDB 및 Cassandra와 같은 오픈 API를 지원하는 관리형 멀티모달 NoSQL 데이터베이스입니다. 이 솔루션은 단일 자릿수 밀리초 응답 시간과 고가용성이 필요한 애플리케이션에 Azure Cosmos DB를 사용합니다. Azure Cosmos DB는 모든 Azure 지역에서 다중 지역 쓰기를 제공합니다. Azure Synapse Link for Azure Cosmos DB를 사용하여 실시간으로 데이터에 대한 통찰력을 얻고 분석을 실행할 수 있습니다.
- Azure Cognitive Search 는 애플리케이션과 API에 필요한 데이터를 인덱싱할 수 있는 클라우드 검색 서비스입니다. Cognitive Search에는 텍스트 추출을 돕고 텍스트가 아닌 파일에서 텍스트를 유추하는 선택적 AI 강화 기능이 있습니다. Cognitive Search

는 Azure Data Lake Storage 및 Azure Cosmos DB와 같은 서비스와 통합되어 데이터에 쉽게 액세스하고 인덱싱합니다. REST API 또는 .NET SDK를 사용하여 인덱싱된 데이터를 쿼리할 수 있습니다. 두 개의 별도 인덱스에서 데이터를 가져오려면 이를 단일 인덱스로 결합하거나 복잡한 데이터 형식을 사용할 수 있습니다 .

시나리오 세부 정보

거의 실시간으로 변경 사항을 처리하기 위한 종단 간 워크플로에는 다음이 필요합니다.

- 변경 데이터 캡처(CDC) 기술. OLTP 애플리케이션은 SQL Server, MySQL, Oracle과 같은 서로 다른 백엔드 데이터 저장소를 가질 수 있습니다. 첫 번째 단계는 변경 사항이 발생하는 대로 이를 수신하고 이를 전파하는 것입니다.
- 변경 이벤트를 대규모로 게시하기 위한 수집 버퍼. 이 서비스는 메시지가 도착하면 대량의 데이터를 처리할 수 있어야 합니다. 개별 구독자는 이 시스템에 연결하여 데이터를 처리할 수 있습니다.
- 원시 형식 그대로의 데이터를 분산 및 확장 가능한 방식으로 저장합니다.
- 사용자가 상태를 다시 시작하고 관리할 수 있는 분산형 효율적인 스트림 처리 시스템입니다.
- 비즈니스 의사결정을 뒷받침하기 위해 대규모로 실행되는 분석 시스템입니다.
- 셀프 서비스 분석 인터페이스.
- 저지연 API 응답의 경우, 데이터의 비정규화된 표현을 저장하는 NoSQL 데이터베이스가 필요합니다.
- 어떤 경우에는 데이터를 색인하고, 정기적으로 색인을 새로 고치고, 최신 데이터를 다운스트림에서 사용할 수 있도록 하는 시스템이 필요합니다.

앞서 언급한 모든 기술은 경계 보안, 인증, 권한 부여 및 데이터 암호화를 위한 관련 보안 구조를 사용해야 합니다.

잠재적 사용 사례

이 솔루션은 다음에 적합합니다:

- OLTP에서 OLAP(온라인 분석 처리)로 변경 사항을 전파해야 하는 산업.
- 데이터 변환이나 보강이 필요한 애플리케이션.

실시간 데이터 처리 시나리오는 금융 서비스 산업에 특히 중요합니다. 예를 들어, 보험, 신용카드 또는 은행 고객이 지불을 한 다음 즉시 고객 서비스에 연락하면 고객 지원 담당자는 최신 정보를 가지고 있어야 합니다.

유사한 시나리오가 소매, 상거래 및 의료 부문에도 적용됩니다. 이러한 시나리오를 활성화하면 운영이 간소화되어 조직 생산성이 높아지고 고객 만족도가 향상됩니다.