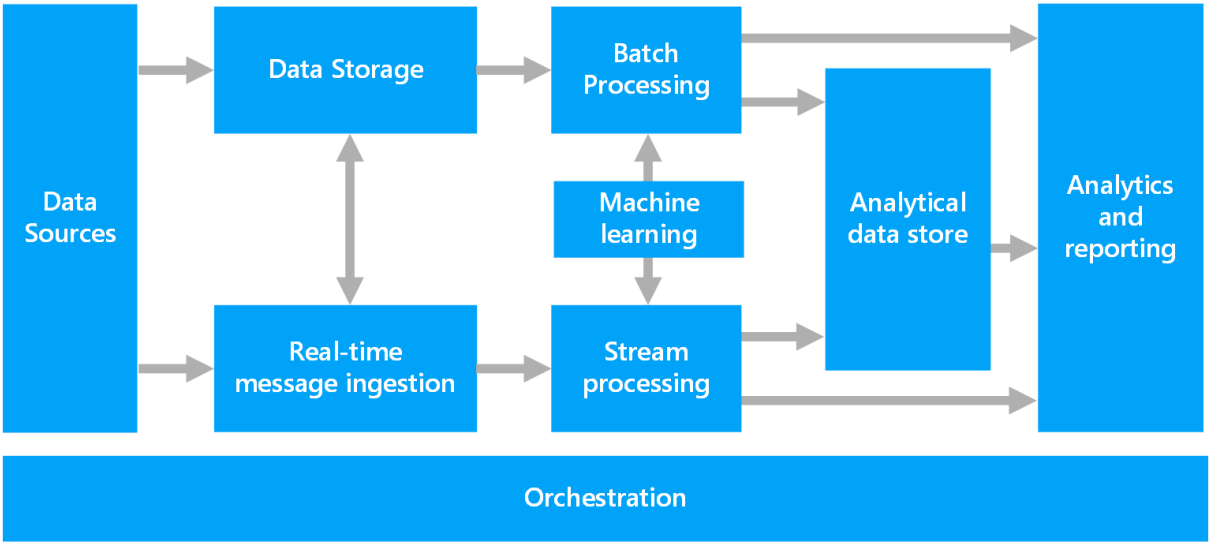
MS-기본/강의요약

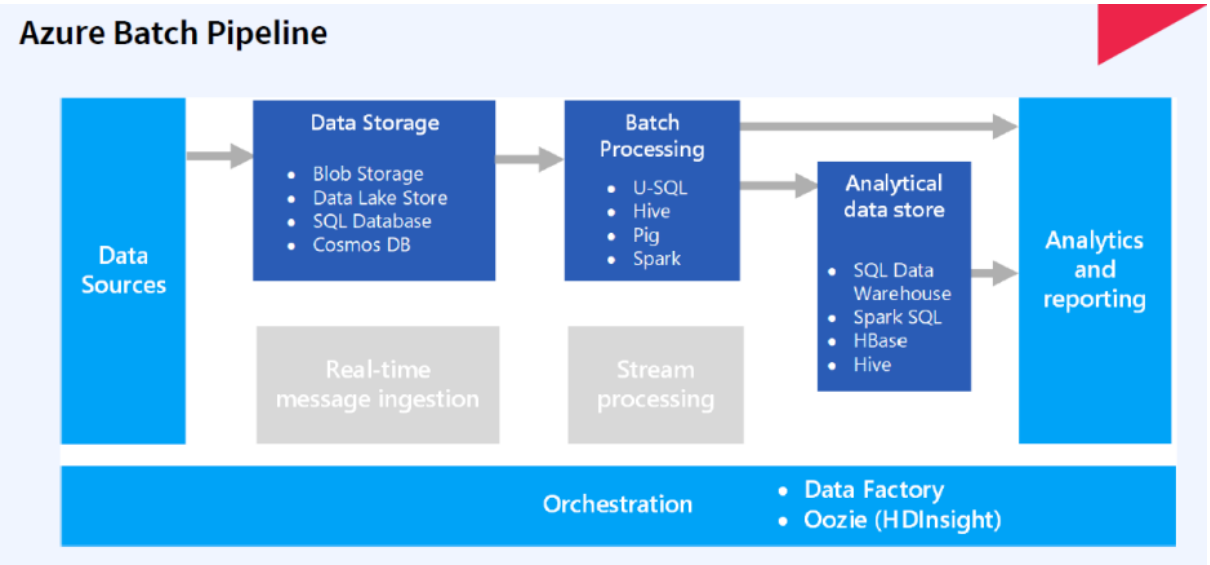
Azure Big Data Architecture

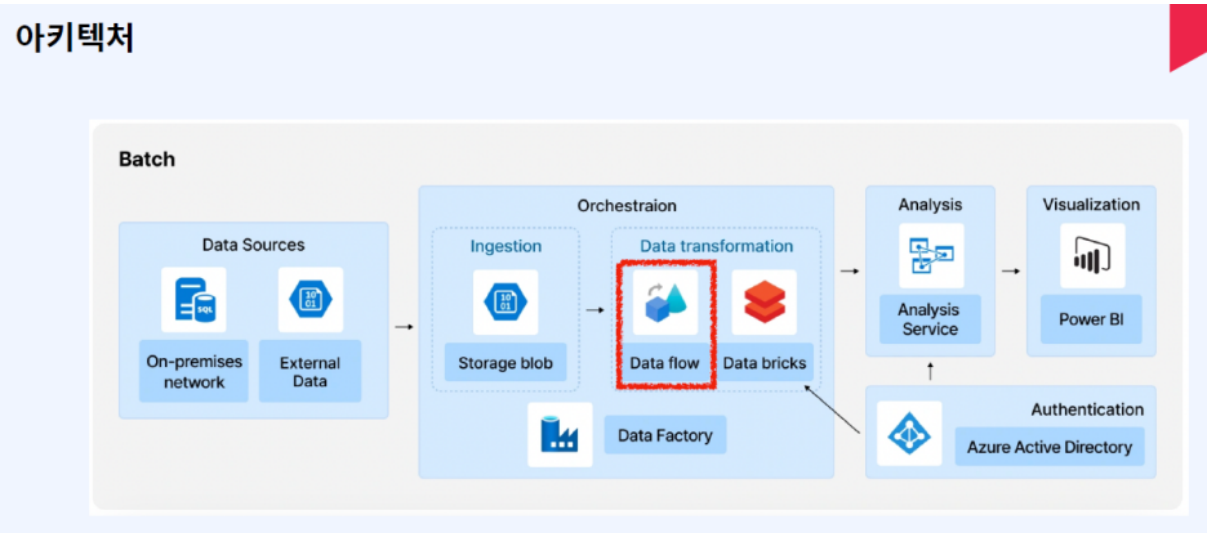


강의 목표

1. Azure 솔루션들을 사용하여 기본적인 데이터 파이프라인을 설계하고 구축할 수 있다.
2. 데이터 일괄 처리와 실시간 처리의 차이점을 이해하고,
   요구사항에 알맞은 처리 방법을 선택하고 해당하는 Azure 솔루션을 이용할 수 있다.
3. 데이터 크기에 따라 알맞은 Azure 솔루션을 선택하여,
   확장성 있고 안정적인 데이터 파이프라인을 구축할 수 있다.
4. 데이터 처리를 할 때, 각각 알맞은 솔루션을 선택하여
   코딩 없이 간단한 처리를 할 수도 있고, 코드를 사용한 복잡한 처리도 할 수 있다.
5. 다른 클라우드 솔루션과 연동된 데이터 파이프라인(멀티 클라우드)을 구축할 수 있다.

배치 파이프라인





1)Data Storage

Azure Blob Containers
. Azure Data Lake Store

Blob Storage
. 오브젝트 스토리지 (Object Storage)
. 대량의 비정형 데이터를 저장하는 데 최적화 (Text, Binary Data, 이미지, 비디오...)
. 계층적 구조 (Multiple Tiers)

Data Lake Gen2
. Azure Blob Storage를 기준으로 하는 빅데이터 분석 전용
. Data Lake Storage Gen1의 기능을 Azure Blob Storage와 통합한 것
. Hadoop 호환 액세스 (기존 Data Lake Storage Gen1 기능)
. 파일 시스템의 의미 체계, 파일 수준 보안 및 확장 제공 (기존 Azure Blob Storage 기능)
→ 고가용성/재해복구 기능

2)Data Process

. Azure Synapse Analytics
. Azure Data Lake Analytics
. HDInsight (Hive, Pig, Spark...)
. Azure Databricks (Spark as a Service)

3) Data Store: Data Warehouse

. Azure Synapse Analytics
. Azure Synapse Spark pools
. Azure Databricks
. Azure Data Explorer
. Azure SQL Database
. SQL Server in Azure VM
. HBase on HDInsight
. Hive LLAP on HDInsight
. Azure Analytical Services
. Azure Cosmos DB

4)Analytics and Reporting

. Azure Analysis Services
. Power BI
. Microsoft Excel

Azure DataFactory
. Data Integration Service (데이터 통합 서비스) - Orchestration Tool, 현대화된 SSIS
. 90개 이상의 커넥터를 추가 비용없이 사용하여 데이터 원본을 가시적으로 통합 가능
→ On-Premise 데이터 및 SaaS 데이터 수집 가능
. 직관적인 환경 내에서 ETL 및 ELT 프로세스를

코딩 없이 손쉽게 생성할 수도 있고(쉬운 사용성), 코드를 직접 작성할 수도 있다
. Azure Synapse Analytics와 연동하여 비즈니스 인사이트를 활용할 수 있다.

Data Process
참고:
https://learn.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing
. Azure Synapse Analytics
. Azure Data Lake Analytics
. HDInsight (Hive, Pig, Spark...)
. Azure Databricks (Spark as a Service)

Azure DataFlow
참고:
https://learn.microsoft.com/ko-kr/azure/data-factory/concepts-data-flow-overview
. GUI 환경에서 할 수 있는 데이터 변환
. 코드를 작성하지 않고도 데이터 변환 로직을 개발할 수 있다
. Scale out Apache Spark 클러스터를 사용하는 Azure Data Factory 파이프라인 내에서 작업 실행
. Azure Data Factory 일정, 제어, 흐름, 모니터링 기능을 사용해 운용할 수 있다
. 일부 지역에서만 서비스 지원

Azure Synapse Analytics
참고:
https://learn.microsoft.com/ko-kr/azure/synapse-analytics/overview-what-is
. 데이터 웨어하우스와 빅데이터 시스템 전체에서 쉽게 인사이트를 얻을 수 있는 엔터프라이즈 분석 서비스
. SQL 기술(데이터 웨어하우징), Spark 기술(빅데이터), 로그 및 시계열 분석(Data Explorer),
ETL/ELT를 위한 파이프라인, PowerBI, CosmosDB, AzureML과 같은 Azure 서비스와의 긴밀한 결합

범주	기능	Azure Data Factory	Azure Synapse Analytics
통합 런타임	지역 간 통합 런타임 지원(데이터 흐름)	✓	✗
	통합 런타임 공유	*✓ 여러 데이터 팩터리에서 공유 가능*	✗
파이프라인 작업	파워 쿼리 작업 지원	✓	✗
	전역 매개 변수 지원	✓	✗
템플릿 갤러리 및 지식 센터	솔루션 템플릿	*✓ Azure Data Factory 템플릿 갤러리*	*✓ Synapse 작업 영역 지식 센터*
GIT 리포지토리 통합	Git 통합.	✓	✓
Monitoring	데이터 흐름에 대한 Spark 작업 모니터링	✗	*✓ Synapse Spark 풀 활용*

Azure Databricks
참고:
https://learn.microsoft.com/ko-kr/azure/databricks/
참고:
https://learn.microsoft.com/ko-kr/azure/databricks/getting-started/concepts
참고:
https://learn.microsoft.com/ko-kr/azure/databricks/scenarios/what-is-azure-databricks-ws
. Apache Spark 기반으로 하는 분석 플랫폼

HDInsight
참고:
https://learn.microsoft.com/ko-kr/azure/hdinsight/
참고:
https://learn.microsoft.com/ko-kr/azure/hdinsight/hdinsight-overview
. 엔터프라이즈용 클라우드의 관리형 오픈소스 분석 서비스
. Azure 환경에서 Spark, Hive, Kafka, Hadoop 과 같은 오픈소스 프레임워크를 사용할 수 있다

Data Format

반정형 데이터 지원 형식
2-2. Data Ingestion - 참고: Data Format
JSON(Javascript Object 표기법)
. JSON 데이터는 모든 애플리케이션에서 생성하는 것이 가능하다
. 공식 사양이 없으므로 다양한 구현 사이에서 차이가 있을 수 있다. JSON 파서마다 규칙이 다를 수 있다.
Avro
. Apache Hadoop과 함께 사용하기 위해 개발된 오픈소스 데이터 직렬화 및 RPC 프레임워크
. JSON에 정의된 스키마를 활용하여 직렬화된 데이터를 압축된 바이너리 형식으로 생성
. 직렬화(참조 형식의 데이터를 모두 값 형식 데이터로 전환 => 언어에 따라서 텍스트 또는 바이너리 등의 형태가
되어서, 저장하거나 통신 시 파싱이 가능한 유의미한 데이터가 됨)된 데이터를 모든 대상으로 전송할 수 있다.
. 데이터에 스키마가 포함되어 있으므로, 편리하게 역직렬화를 수행할 수 있다
. Avro 스키마는 스키마 타입 및 스키마 타입에 대한 데이터 속성을 정의하는 JSON 문자열, 오브젝트, 배열로 구성

반정형 데이터 지원 형식
2-2. Data Ingestion - 참고: Data Format
ORC(Optimized Row Columnar)
. Hive 데이터를 저장하는 데 사용되는 이진 형식
. ORC는 이전의 Hive 파일 형식보다 데이터 읽기, 쓰기, 처리 성능을 개선하고 효율적으로 압축하기 위해 설계됨
Parquet
. Hadoop 에코 시스템의 프로젝트용으로 설계된 효율적인 압축 열 형식 데이터 표현
. 복잡한 중첩 데이터 구조를 지원
. 컬럼 기반 저장 포맷으로 데이터를 미리 컬럼 단위로 압축시키고, 필요한 칼럼만 읽고, 집계하는 데 빠르다
=> 필요한 데이터만 디스크로부터 읽어 I/O를 최소화하고, 데이터 크기를 줄이기 위한 목적이다.
. 압축률이 좋고, 컬럼 단위로 구성하면 데이터가 균일하므로, 압축률이 높아 파일의 크기도 작다
. 컬럼별로 적합한 인코딩을 할 수 있다.
빠르게 읽고, 압축률이 좋고, 특정 언어에 종속되지 않은 빅데이터 처리에 좋은 포맷:
Parquet, ORC, Avro

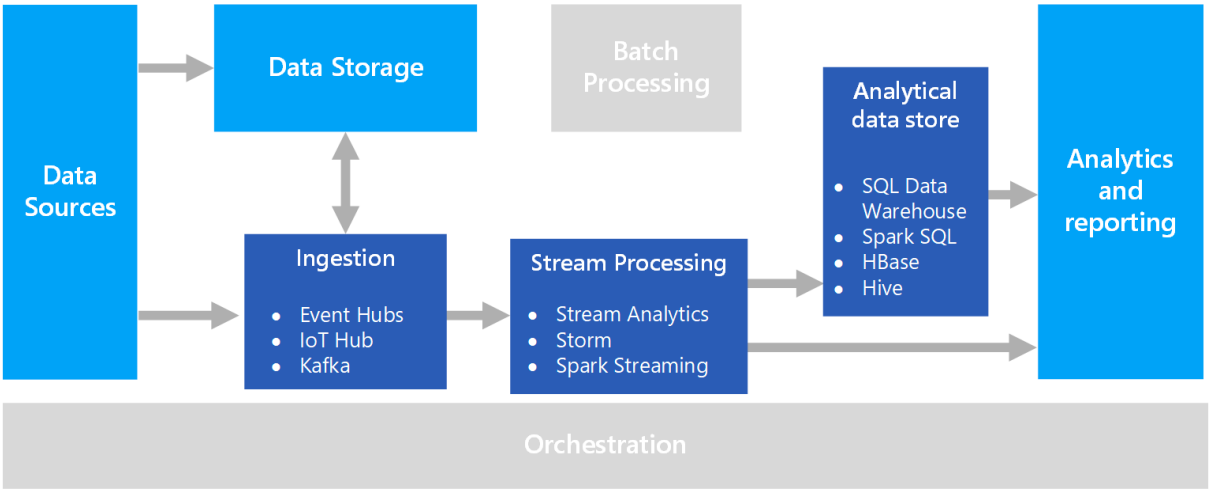
반정형 데이터 지원 형식
2-2. Data Ingestion - 참고: Data Format
XML(eXtensible Markup Language)
. 문서를 인코딩하기 위한 일련의 규칙을 정의하는 마크업 언어
. 문서를 구성하는 구조 및 요소를 표준화하기 위해 개발된 마크업 언어인 SGML을 기반으로 개발
. 초기의 문서 중심에서 확장되어 임의 데이터 구조의 표현 및 통신 프로토콜의 기본 언어 등 광범위한 용도로 사용
. 시작 태그와 일치하는 종료 태그로 구성

참고: Data 압축
데이터 압축은 원래 데이터를 더 적은 비트로 인코딩하는 과정. 데이터 압축은 대역폭과 I/O를 줄이고,
스토리지에서 랜덤 액세스 메모리 및 디스크 공간을 절약하는데 유용하다.
압축 알고리즘 종류

1. 무손실 압축: 원래 데이터가 압축 해제 될 때 완전히 복구되는 방식
   => 저장된 데이터가 약간 변경되는 것만으로도 사용할 수 없을 우려가 있기 때문에 대부분 많이 쓰임
2. 손실 압축: 비디오, 오디오 및 이미지 압축에 유용
   데이터 인코딩:
   사람이 인지할 수 있는 형태의 데이터를 약속된 규칙에 의해 컴퓨터가 이해할 수 있는 0과 1로 변환하는 과정

실시간 파이프라인 Azure Real-time Pipeline

데이터 스트림 사용. 새로운 동적 데이터 시계열로 생성 시 유리, 최소대기시간



센서 데이터

시간 경과에 따른 데이터 집계

새로운 동적 데이터 생성

주식시간 변동.

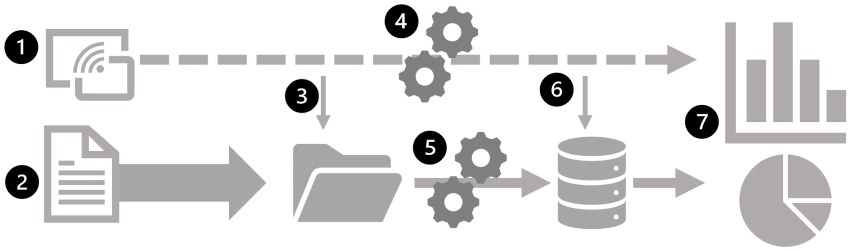
메시지를 실시간 - 수집 파이프라인 차단하지 않으면서 수행.

대용량 지원필요

수집 파이프라인 차단

일괄 처리, 실시간 처리 혼용 카파

일괄 + 스트림 처리



원본 데이터 캡쳐

임시 윈도우에서 데이터 필터링 또는 집계

메시지 수집 및 브로커

(1) 실시간 메시지 수집

Azure Solutions
• Azure Event Hubs: 초당 수백만 개의 이벤트 메시지를 수집하기 위한 메시징 솔루션
• Azure IoT Hub: 인터넷에 연결된 디바이스 간의 양방향 통신과 동시에 연결된
수백만 대의 디바이스를 처리할 수 있는 확장 가능한 메시지 큐를 제공

• Apache Kafka: 다중 메시지 생산자에서 들어온 초당 수백만 개의 메시지를 처리하여 여러 소비자에게 전송하기 위해 확장될 수 있는 오픈 소스 메시지 큐 및 스트림 처리 애플리케이션

(2) 데이터 스토리지

- Azure Storage Blob 컨테이너 또는 Azure Data Lake Store
  • 들어오는 실시간 데이터는 일반적으로 메시지 브로커에서 캡처되지만,
  일부 시나리오에서는 폴더에 새 파일이 있는지 모니터링한 후 생성 또는 업데이트될 때 처리하는 것이 적절할 수 있습니다.
  • 많은 실시간 처리 솔루션은 스트리밍 데이터를 파일 저장소에 저장될 수 있는 정적 참조 데이터에 결합
  • 파일 스토리지를 보관하거나 람다 아키텍처에서 추가로 일괄 처리하기 위해
  캡처된 실시간 데이터에 대한 출력 대상으로 사용

(3) 스트림 처리

Azure Solutions
• Azure Stream Analytics: 바인딩되지 않은 데이터 스트림에 대해 영구 쿼리를 실행
• Storm: Apache Storm은 Spout 및 Bolt 토폴로지를 사용
실시간 처리 데이터 원본의 결과를 사용, 처리 및 출력하는 스트림 처리를 위한 오픈 소스 프레임워크
• Spark Streaming: Apache Spark는 일반 데이터 처리를 위한 오픈 소스 분산 플랫폼

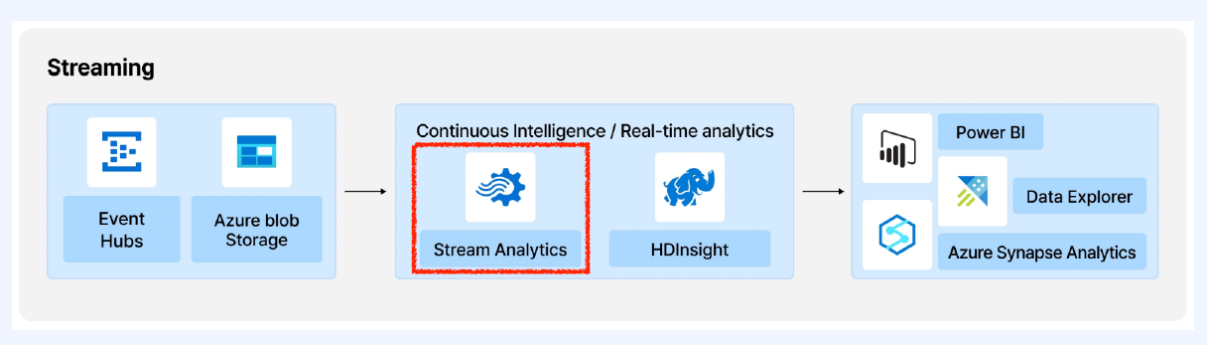
(4) 분석 데이터 저장소

- Azure Synapse Analytics, Azure Data Explorer, HBase, Spark 또는 Hive:
  처리된 실시간 데이터는 Synapse Analytics, Azure Data Explorer와 같은 관계형 데이터베이스 or
  HBase와 같은 NoSQL 스토리지 or
  Spark 또는 Hive 테이블을 정의하고 쿼리할 수 있는 분산 스토리지에 파일로 저장

(5) 분석 및 보고

Azure Analysis Services, Power BI 및 Microsoft Excel:
• 분석 데이터 저장소에 저장된 처리된 실시간 데이터 분석 데이터는
일괄 처리된 데이터와 같은 방식으로 기록 보고 및 분석에 사용
• Power BI는 대기 시간이 충분히 낮은 분석 데이터 원본에서 또는
경우에 따라 스트림 처리 출력에서 직접
실시간(또는 거의 실시간) 보고서 및 시각화를 게시하는 데 사용

전체 파이프라인 개요 - 아키텍쳐



Azure Stream Analytics

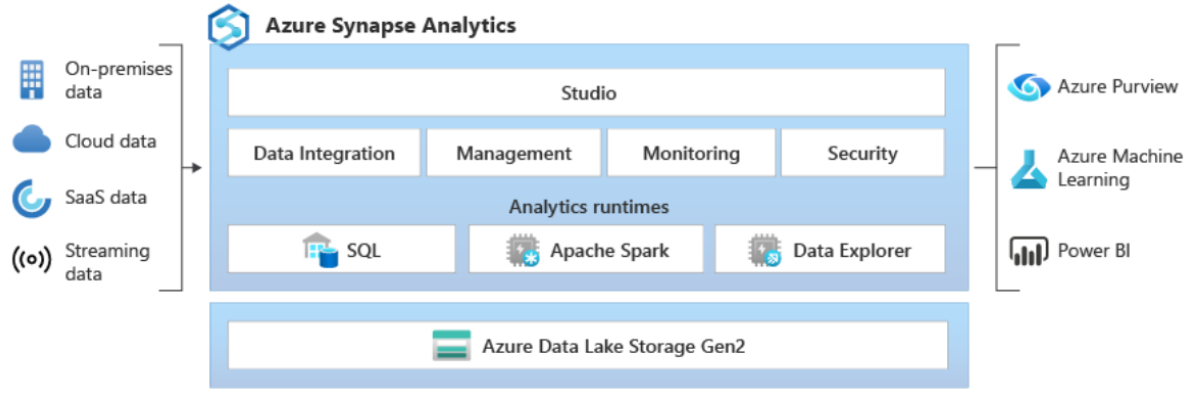
다양한 원본에서 실시간 데이터 스트림을 필터링, 집계 및 처리하는 데 사용할 수 있는 클라우드 기반 스트림 처리 엔진을 제공

• 처리 결과를 사용하여 서비스 또는 애플리케이션에서 자동화된 작업을 트리거하거나, 실시간 시각화를 생성하거나, 스트리밍 데이터를 엔터프라이즈 분석 솔루션에 통합

사례 : 클릭스트림 기반 추천 , Iot 센서 ,금융 이상거래 탐지

대기시간이 밀리초 미만인 대용량 스트리밍 데이터를 분석 처리하도록 설계된 완전 관리 스트림 처리 엔진.

Azure Synapse Analytics



- 엔터프라이즈 데이터 웨어하우징과 빅 데이터 분석을 결합한 무제한 분석 서비스
  • Azure Stream Analytics 를 통해 Azure Synapse Analytics의 전용 SQL 풀 테이블로 출력
  => 최대 200MB/초의 처리량 속도
  • 리포트 및 대시보드와 같은 워크로드에 대한 가장 까다로운 실시간 분석 및 실행 부하 과다 경로 데이터 처리 요구 사항을 지원

엔터프라이즈 데이터웨어하우징과 빅데이터 쿼리 분석을 결합한 무제한 분석 서비스

전용 SQL 풀 테이블로 출력할 수 있으며 최대 200MB/초 처리량 속도를 처리

HD Insight

Azure환경에서 Apache Spark, Hive, LLAP, Kafka, Hadoop 등을 사용하여 대용량 및 빠른 속도로 빅데이터 프레임 워크를 실행하는 것을 간소화하는 전체 스펙트럼 관리형 클러스터 플랫폼.

오픈소스의 솔루션 최소 성능 보장 MS 에서

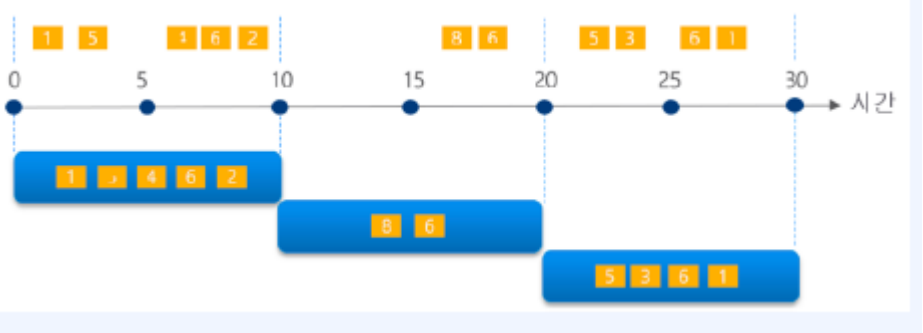
분산시스템 고가

자체 온프레미스로 구축

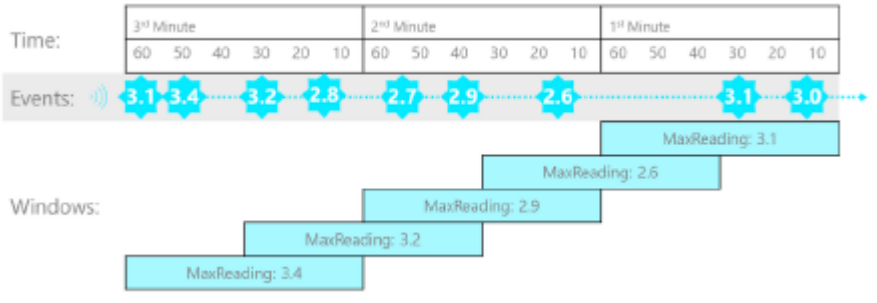
실시간 처리 실습

스트림 처리의 일반적인 목표는 이벤트를 임시 간격 또는 창으로 집계하는 것입니다.

1. 연속: 데이터 스트림을 고정된 크기의 겹치지 않는 시계열 창으로 나누고, 해당 시계열에 대해 작동
   => 이벤트는 둘 이상의 연속 window에 속하지 않음



1. 도약: 예정된 겹치는 window를 모델링하고 정해진 기간만큼 시간을 앞으로 이동
   => 이벤트는 두 개 이상의 window 결과 집합에 속할 수 있음



슬라이딩

모든 창에 최소 1개

세션

스냅샷

슬라이딩 윈도우와 텀블링 윈도우 비교

슬라이딩 윈도우와 **텀블링 윈도우**는 실시간 데이터 처리나 시계열 데이터 분석에서 자주 사용되는 개념입니다. 둘 다 일정한 크기의 데이터 조각을 추출하여 분석하는데 사용되지만, 윈도우가 이동하는 방식에 따라 차이가 있습니다.

슬라이딩 윈도우 (Sliding Window)

- **정의:** 고정된 크기의 윈도우가 데이터 스트림을 따라 미끄러져 이동하며, 윈도우 내의 데이터를 분석하는 방식입니다.
- **특징:**
  - 윈도우가 일정하게 겹치면서 이동합니다.
  - 윈도우의 시작과 끝이 시간에 따라 지속적으로 변합니다.
  - 데이터의 변화를 실시간으로 반영하고, 더욱 부드러운 추세를 파악하는 데 유용합니다.
  - 이동 평균, 이동 표준편차 등을 계산하는 데 주로 사용됩니다.

예시:

- 주식 시장 데이터에서 5분 간격으로 슬라이딩 윈도우를 설정하여 이동 평균을 계산하면, 주가의 단기적인 변동을 더욱 정확하게 파악할 수 있습니다.
- IoT 센서 데이터에서 1초 간격으로 슬라이딩 윈도우를 설정하여 실시간으로 온도 변화를 모니터링할 수 있습니다.

텀블링 윈도우 (Tumbling Window)

- **정의:** 고정된 크기의 윈도우가 시간에 따라 일정한 간격으로 새롭게 생성되면서 이전 윈도우와 겹치지 않는 방식입니다.

- **특징:**
  - 윈도우가 서로 겹치지 않고 연속적으로 생성됩니다.
  - 각 윈도우는 독립적인 데이터 집합으로 간주됩니다.
  - 특정 시간 간격 동안의 데이터를 요약하고 집계하는 데 주로 사용됩니다.

예시:

- 웹 서버 로그에서 1시간 간격으로 텀블링 윈도우를 설정하여 시간대별 페이지 조회수를 집계할 수 있습니다.
- IoT 센서 데이터에서 1일 간격으로 텀블링 윈도우를 설정하여 일별 평균 온도를 계산할 수 있습니다.

슬라이딩 윈도우와 텀블링 윈도우의 차이점 요약

특징	슬라이딩 윈도우	텀블링 윈도우
윈도우 이동 방식	겹치면서 이동	겹치지 않고 새로 생성
데이터 분석 목적	실시간 변화 파악, 부드러운 추세 분석	특정 시간 간격의 요약 및 집계
주요 활용 분야	이동 평균, 이동 표준편차 등	시간별 통계, 배치 처리

Sheets로 내보내기

어떤 윈도우를 사용해야 할까요?

- **슬라이딩 윈도우:** 실시간 데이터 분석, 빠른 변화 감지, 예측 모델 학습 등에 적합합니다.
- **텀블링 윈도우:** 배치 처리, 보고서 생성, 오프라인 분석 등에 적합합니다.

선택 기준:

- **데이터의 특성:** 데이터가 얼마나 빠르게 변화하는지, 어떤 주기로 분석해야 하는지에 따라 선택합니다.
- **분석 목표:** 실시간 분석이 필요한지, 특정 시간 간격의 요약이 필요한지에 따라 선택합니다.
- **시스템 자원:** 슬라이딩 윈도우는 더 많은 계산량을 요구하므로 시스템 자원을 충분히 고려해야 합니다.

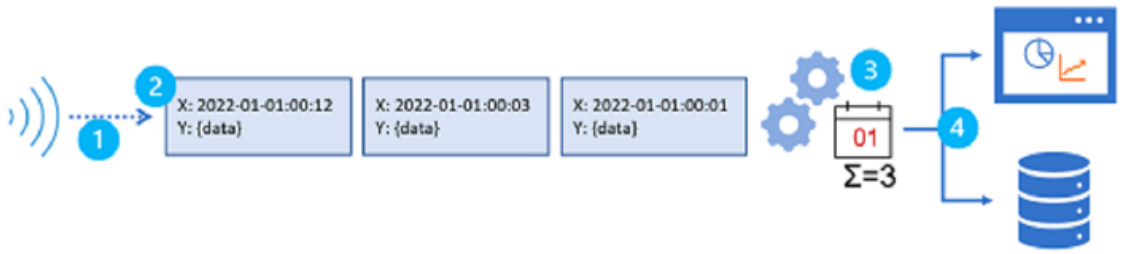
결론적으로, 슬라이딩 윈도우와 텀블링 윈도우는 각각 장단점이 있으며, 데이터 분석 목적과 시스템 환경에 맞게 적절한 윈도우를 선택해야 합니다.

실시간 데이터 수집 방법 소개

• Stream Analytics는
정확히 한 번에 이벤트를 처리하고 한 번 이상 이벤트가 전송되도록 보장 => 이벤트 손실 X
이벤트 전송에 실패하는 경우에 기본 제공 복구 기능

• Stream Analytics는 작업의 상태를 유지하는 기본 검사점을 제공하고 반복 가능한 결과를 생성

=⇒ 데이터 정합 보장.



실시간 데이터 저장 방법 소개: DeltaLake (Delta Table)
• Databricks Lakehouse 플랫폼에서 데이터 및 테이블을 저장하기 위한 기반을 제공하는 최적화된 스토리지 계층
• ACID 트랜잭션 기능 제공
• 확장 가능한 메타데이터 처리를 위한 파일 기반 트랜잭션 로그로 Parquet 데이터 파일을 확장하는 OSS
• Apache Spark API와 완벽하게 호환되며 Structured Streaming과의 긴밀한 통합을 위해 개발됨
• 일괄 처리 및 스트리밍 작업 모두에 단일 데이터 복사본을 쉽게 사용할 수 있음
• 대규모 증분 처리 제공

- Azure Synapse Data Explorer는 로그 및 원격 분석 데이터에서 인사이트를 잠금 해제할 수 있는
  대화형 쿼리 환경을 고객에게 제공합니다.
  • 최적화된 언어인 KQL(Kusto Query Language) 데이터 탐색기 테이블 쿼리 언어.
  특히 타임스탬프 특성이 포함된 원격 분석 데이터에서 빠른 읽기 성능

Azure Data Catalog
데이터 카탈로그는 사용자가 필요한 데이터 원본을 검색하고 찾은 데이터 원본을 이해할 수 있게 해주는
완전히 관리되는 클라우드 서비스입니다 동시에 Data Catalog 는 조직이 기존 투자에서 더 많은 가치를 얻을 수 있도록 지원합니다
효율적인
Data Catalogue 6 가지 핵심 요소

1. 커넥터 및 큐레이션 도구
2. 자동화
3. 효율적인 검색 옵션
4. 계보 또는 수명 주기 추적
5. 유니버설 용어집 및 데이터 사전
6. 프로파일링

Azure
데이터 카탈로그는 사용자가 필요한 데이터 원본을 검색하고 찾은 데이터 원본을 이해할 수 있게 해주는
완전히 관리되는 클라우드 서비스입니다 동시에 Data Catalog 는 조직이 기존 투자에서 더 많은 가치를 얻을 수 있도록 지원합니다

Azure Data Lake Analytics
는 빅 데이터를 간소화하는 주문형 분석 작업 서비스입니다 .
하드웨어를 배포 , 구성 및 조정하는 대신 , 데이터를 변형하고 귀중한 통찰력을 얻기 위한 쿼리를 작성합니다
이 분석 서비스는 필요한 전력량을 다이얼로 설정하여 어떤 규모의 작업도 즉시 처리할 수 있습니다 .
실행 중일 때만 작업에 대한 비용을 지불하므로 비용 효율적입니다
Azure Data Lake Analytics 7 가지 특징

1. 동적 크기 조정
2. 친숙한 도구를 사용하여 더 빠르게 개발하고 , 디버그하고 , 더 스마트하게 최적화
3. U SQL: 간편하고 , 친숙하고 , 강력한 확장성
4. IT 투자와 완벽하게 통합
5. 저렴하고 비용 효율적
6. 모든 Azure 데이터를 사용하여 작업
7. 지역 내 데이터 보존