



[Snowflake] 1-1. Overview

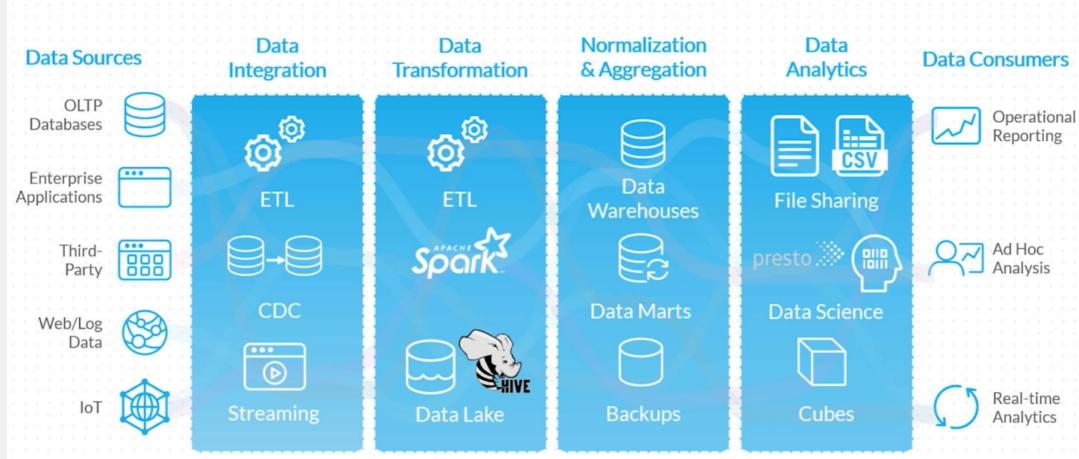


[노션 웹 공유 링크](#) (댓글 & 상세설명 참고)

References

- [데이터 엔지니어링과 Snowflake](#)
- [Snowflake Learn \(SnowPro PREP-CORE Course\)](#). 1장 1강
- [Snowflake 설명서](#)

데이터 엔지니어링



- 기존 데이터 엔지니어링 아키텍처 환경은 Source → Consumer로 이어지는 과정
 - 데이터 소스 → 데이터 통합 → 변형 → 정규화 & 집합 → 분석 → 데이터 소비
- 데이터 엔지니어들은 데이터를 수집하고 분석 할 수 있는 거대한 파이프라인 시스템을 구축하는 역할
- 기업 데이터를 잘 활용하도록 관리하기 때문에 중요한 필수 직군으로 자리매김



ETL 과 ELT, 데이터 웨어하우스 (DW)

- **ETL**
 - Extract, Transform, Load
 - 서로 다른 RDBMS에서 데이터를 추출한 후 변환하고 데이터 웨어하우스에 적재
- **ELT**
 - 추출해서 먼저 적재 후 적재한 곳에서 처리하는 방식을 말한다.
 - 보통 대용량 처리에서 사용
- **데이터 웨어하우스**
 - 데이터 분석, 데이터 마이닝, 인공지능 및 머신러닝을 지원하기 위해 서로 다른 소스의 데이터를 하나의 데이터 저장소로 집계하는 시스템

❓ OLTP(Online Transaction Processing) vs OLAP(Online Analytical Processing)

- OLTP

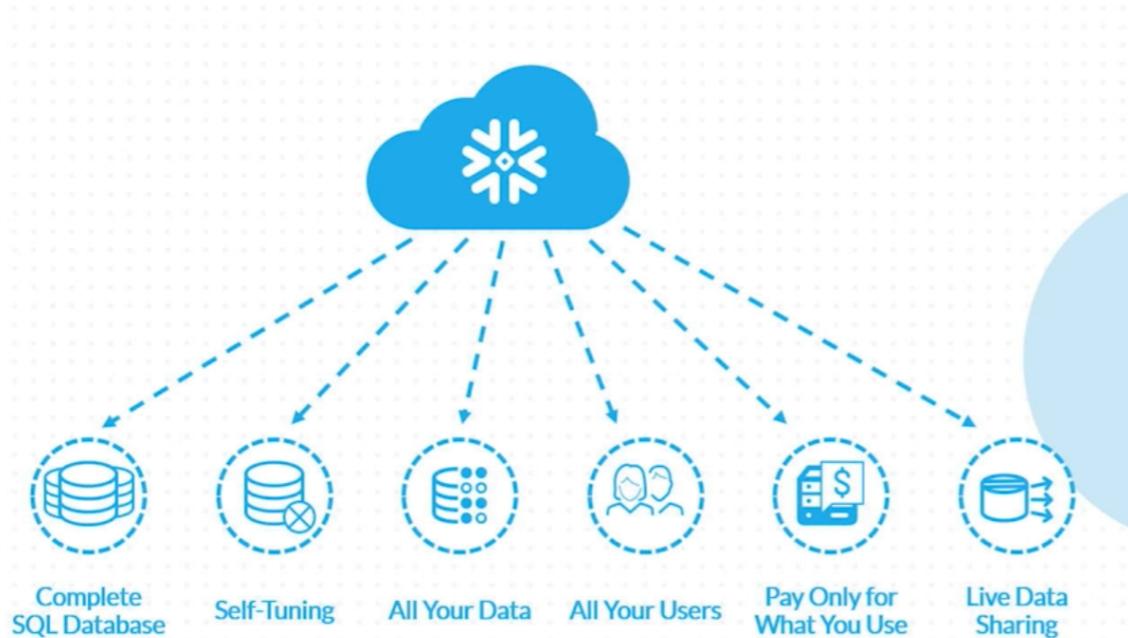
- 현재 현황을 파악하기 위한 목적으로 사용
- 트랜잭션의 효율적인 처리와 안전한 복구를 처리할 수 있도록 해주는 시스템
- 데이터의 조회 및 입력, 수정, 삭제 등의 용도로 사용되는 데이터베이스 시스템

- OLAP

- 온라인 분석 처리를 위한 용도로 사용되는 시스템
- 많은 양의 데이터를 분석하여 다양한 기법을 통해 산출 된 정보를 제공하여 의사결정에 활용

→ 데이터웨어하우스란 OLTP, ETL을 통해 축적된 데이터를 통합하여 관리하는 데이터 창고

- Snowflake Overview



- 데이터 저장, 데이터 처리 부터 시각화, 머신러닝까지 한번에 할 수 있는 클라우드 기반 통합 데이터 플랫폼 솔루션
- 주요 6가지 서비스 (워크로드)

데이터 엔지니어링	데이터 엔지니어링 서비스를 통해 다양한 부서에서 SQL을 이용해 데이터 파이프라인을 효율적으로 구축하고 관리
데이터 레이크	모든 유형의 데이터를 보관할 수 있는 대규모 저장창고로 보안기능이 추가 되어 모든

	데이터를 안전하게 저장
데이터 웨어하우스	데이터 웨어하우스를 통해 snowflake에 분석 가능한 형태로 가공된 데이터를 저장
데이터 사이언스	통계 분석 툴, 머신러닝 기능 등을 제공받아 방대한 양의 데이터를 분석, 다양한 프로그래밍 언어를 통한 데이터 분석
데이터 어플리케이션	데이터 분석 어플리케이션을 신규로 개발 가능하며, 기존 어플리케이션을 Snowflake 와 연동 가능
데이터 교환	데이터를 공유하고 서로 연결하고 협업할 수 있는 솔루션을 제공, 데이터 허브 역할로 정보교환 및 협력기업들도 빠른 데이터 교환이 가능

- **Snowflake 의 주요 특징**

- ANSI SQL에 준거한 **Cloud Data Platform 솔루션**
- 모든 정형, 비 정형 데이터를 고객이 간단하게 로드하여 쿼리
- 모든 시점의 데이터에 대해 실행할 수 있어 분석과 쿼리
- 자가 튜닝, 복구 시스템으로 파티셔닝, 베倜(VACUUM) 동작, 하드웨어 장애 대처가 필요 없음
- AWS (Amazon Web Services), MS Azure, GCP (Google Cloud Platform) 에서 서비스로 제공
- 구독(Subscription)형 라이선스 모델 (하드웨어 구성 및 소프트웨어 라이센스가 없음)
- 일대일 또는 일대다로의 데이터 공유
- 안전한 데이터 공유 및 특정 데이터 세트 공유 (Public Data Marketplace, 데이터 세트 사용)



파티셔닝(Partitioning) 기법

논리적인 데이터 element들을 다수의 entity로 쪼개는 행위
테이블을 파티션이라는 작은 단위로 나누어 관리
데이터베이스를 분산처리하여 성능 저하 방지 및 관리 수월



베倜(Vacuum) 기법

데이터베이스의 일종의 디스크 조각모음
변경 또는 삭제된 자료들이 차지 하고 있는 디스크 공간(ex. 데드튜플)을 다시 사용하기 위한 디스크 공간 확보 작업
이외에도 통계정보 갱신 및 실자료 지도 갱신 등 다양한 이유로 베倜 기법을 사용(PostgreSQL 문서 참고).

- Snowflake vs 기존 데이터 아키텍처



- 기존 데이터 아키텍처 환경은 복잡하고 비용이 많이 드는 제약이 존재, 하지만 Snowflake는 전혀 필요없는 구조
- 데이터 웨어하우징, 레이크, 엔지니어링, 교환, 어플리케이션 및 사이언스 워크로드를 모두 지원
- Snowflake는 OLAP, 온라인 분석처리 시스템으로 설계
 - OLTP 시스템으로 설계 X, OLTP 솔루션으로 사용하지 않는 것을 권장
- 기존 아키텍처와의 차이 & 주요 특장점



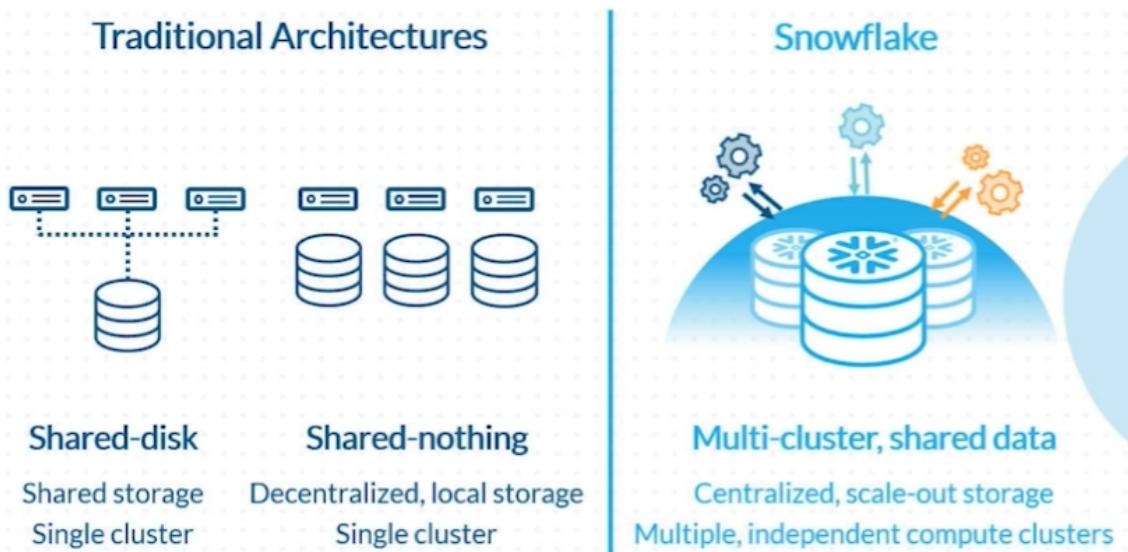
1. 많은 워크로드를 지원

- 데이터 웨어하우징, 데이터 레이크, 데이터엔지니어링, 데이터 교환, 데이터 어플리케이션, 데이터 사이언스뿐만 아닌 비즈니스 부서, 부서 그룹 단위의 워크로드도 지원

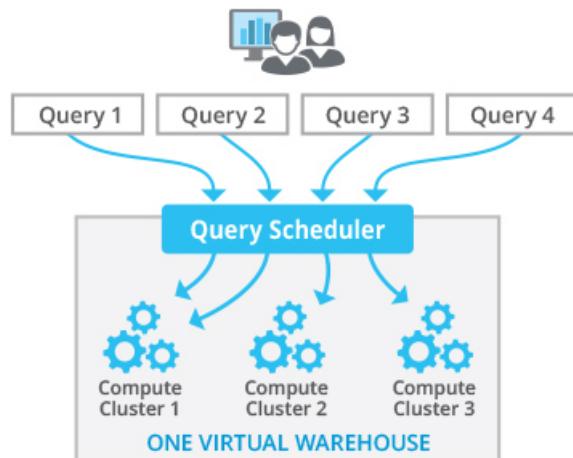
2. 하나의 플랫폼에서 무제한 성능

- 하나의 데이터에 액세스하고 무제한의 성능을 지원하며, 스케일 및 탄력성으로 많은 워크로드와 유저 케이스 지원

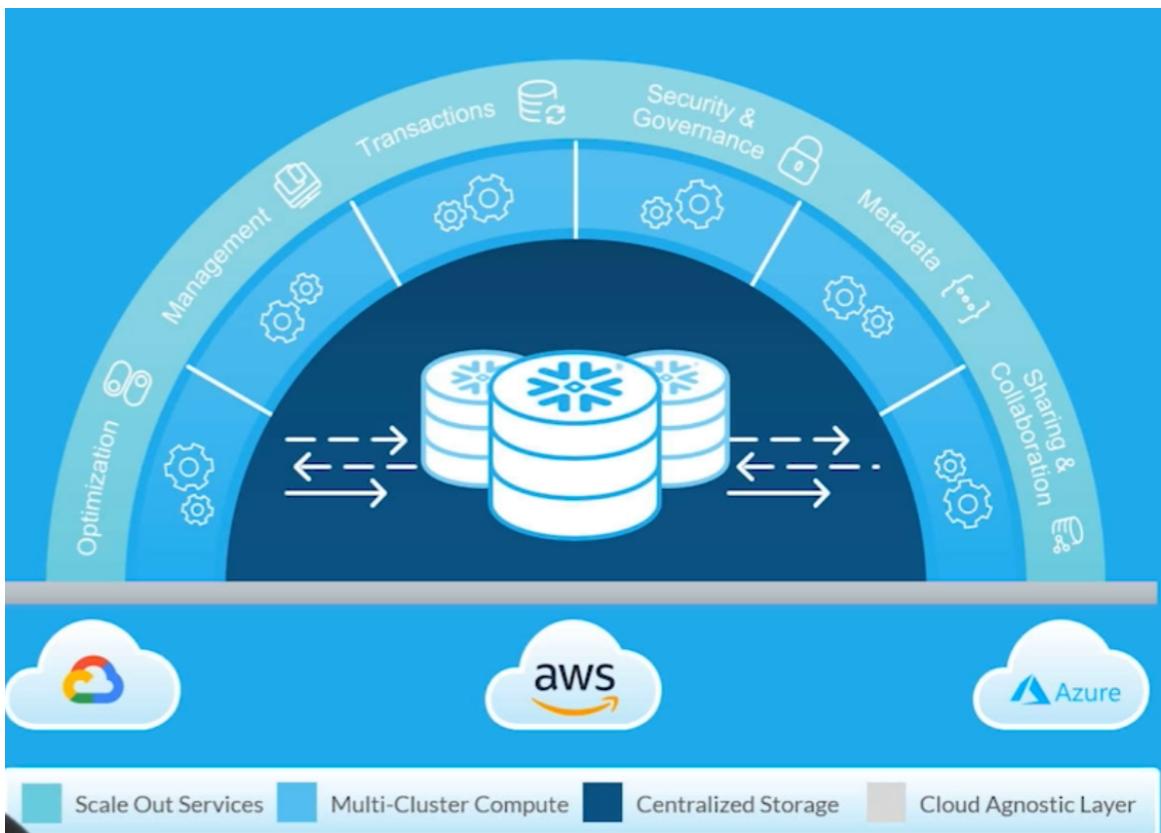
3. 모든 데이터에 대한 액세스는 안전하게 관리
 - 인증된 사용자 및 권한이 있는 사용자만 액세스
4. 서비스로서 Near-Zero 유지보수를 지원
 - 관리팀은 파티셔닝, 베倜작업 등의 일상적인 간단한 작업을 할 필요가 없이 자동으로 실행



5. Multi-cluster, Shared data 아키텍처

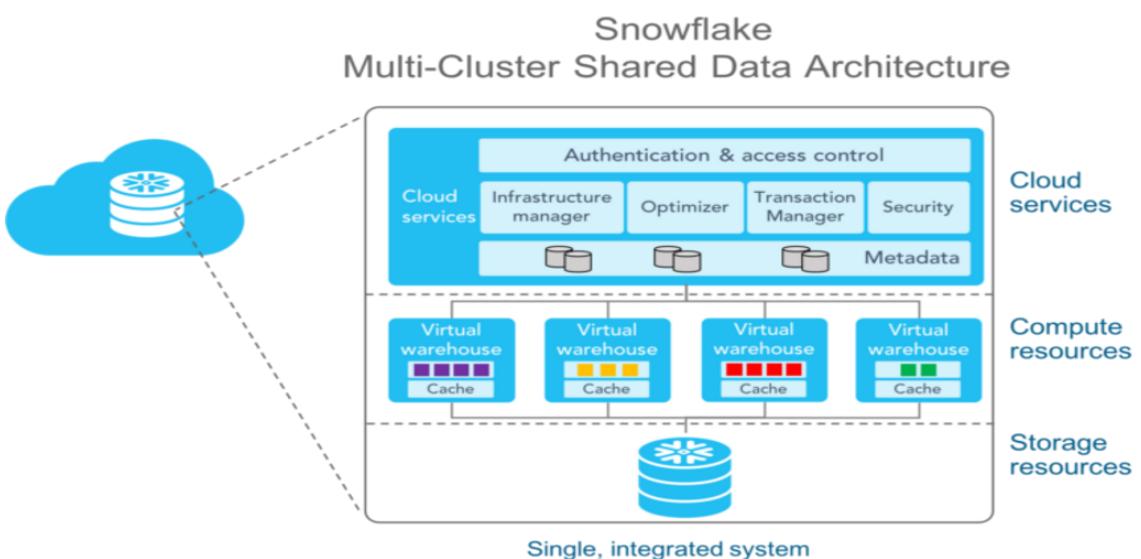


- 멀티클러스터 아키텍처, 멀티클러스터 컴퓨팅, 공유 데이터, 중앙 집중식으로 스토리지 확장
- 거의 무한대의 여러 동립 컴퓨팅 클러스터에서 액세스 및 구성 가능
- **Snowflake 아키텍처 (3-Layer 구조)**

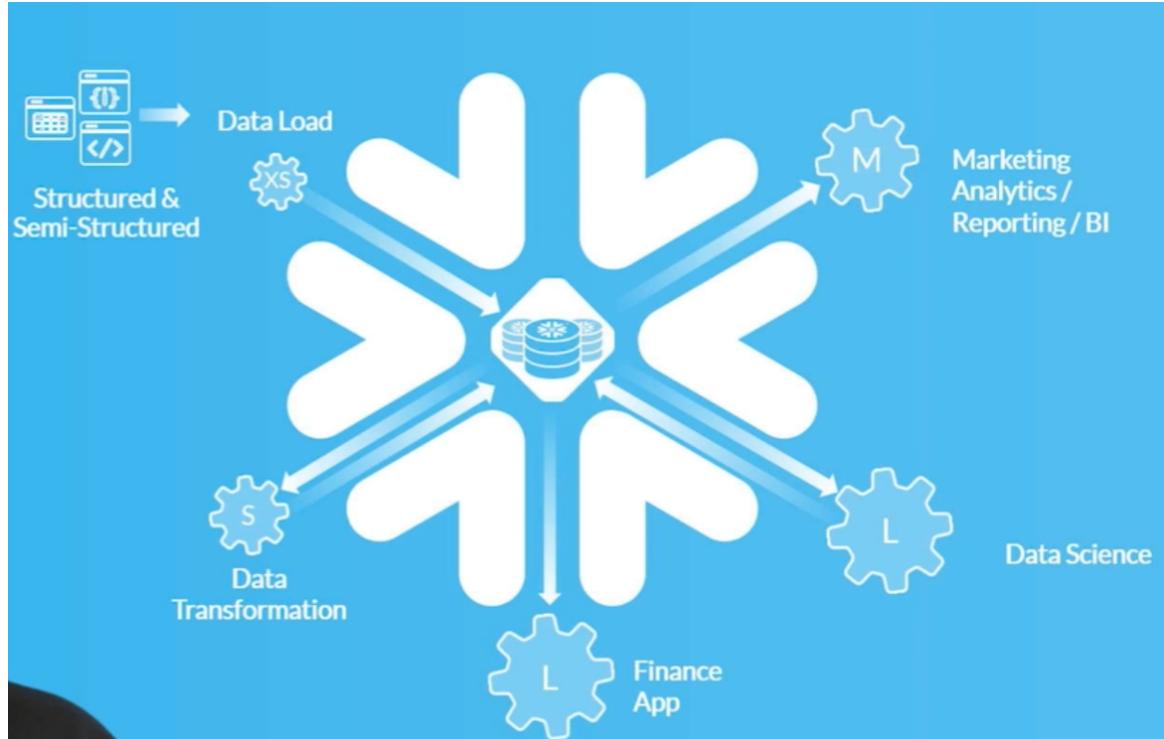


- 클라우드 서비스 레이어
 - 멀티클러스터 컴퓨팅 레이어
 - 중앙 집중식 스토리지 레이어
- 추가적으로 이러한 세개의 레이어는 클라우드 지원 레이어 위에서 동작

- **Multi-Cluster, Shared-Data 아키텍처**



- 클라우드 서비스 레이어, 컴퓨팅 레이어, 스토리지 레이어로 구성된 예시
 - VPC/VNet 과 같이 클라우드 제공업체에서 관리하는 가상 네트워크내에 배포
 - 서비스 레이어에서는 Snowflake 시스템의 엑세스 컨트롤 및 Infrastructure 관리, 옵티마이징, 메타데이터 관리, 보안 등의 역할 수행
 - 가상 웨어하우스(컴퓨팅 레이어)에서는 쿼리문을 실행
 - 스토리지 레이어에 데이터 저장
- 중앙 집중형 스토리지 관리

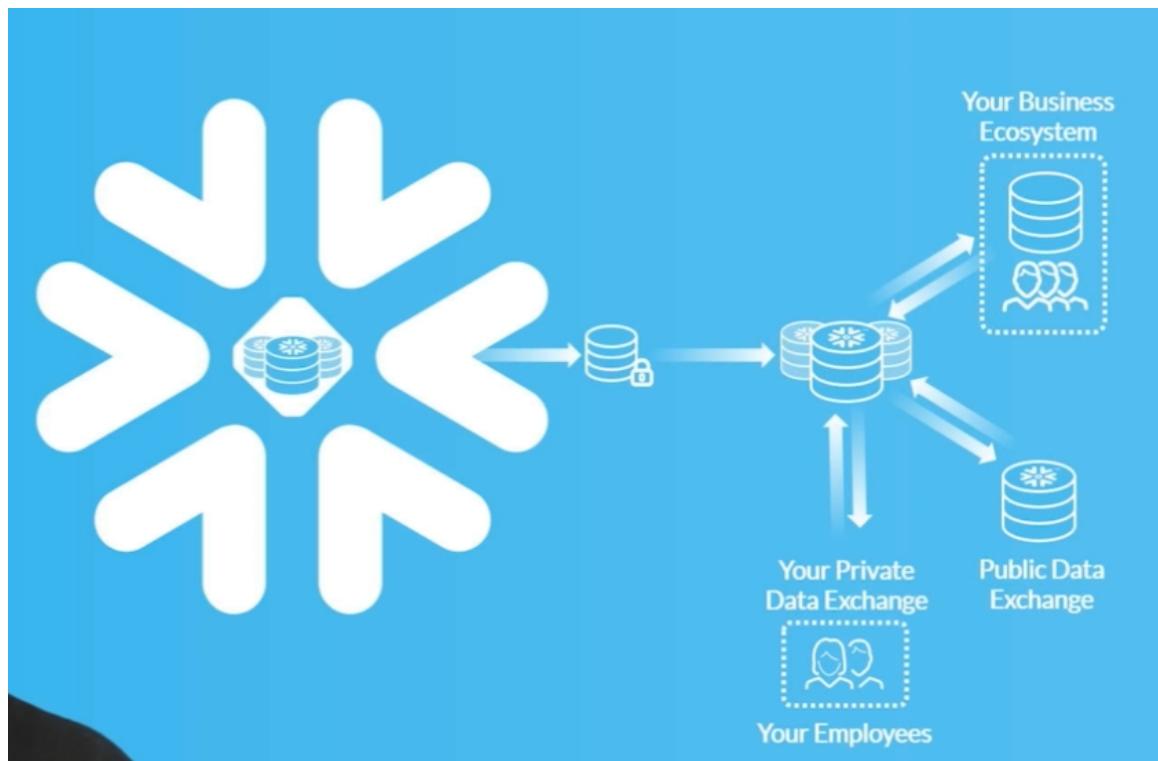


- 각 품니바퀴는 하나의 워크로드이자 Virtual Warehouse (컴퓨팅 레이어)
- 쿼리에 시간이 많이 걸리거나 시간이 충분하지 않을 경우 가상 웨어하우스의 크기를 리사이징 (Scale-up, Down) 하여 해결 가능 (XS ~ 5XL)

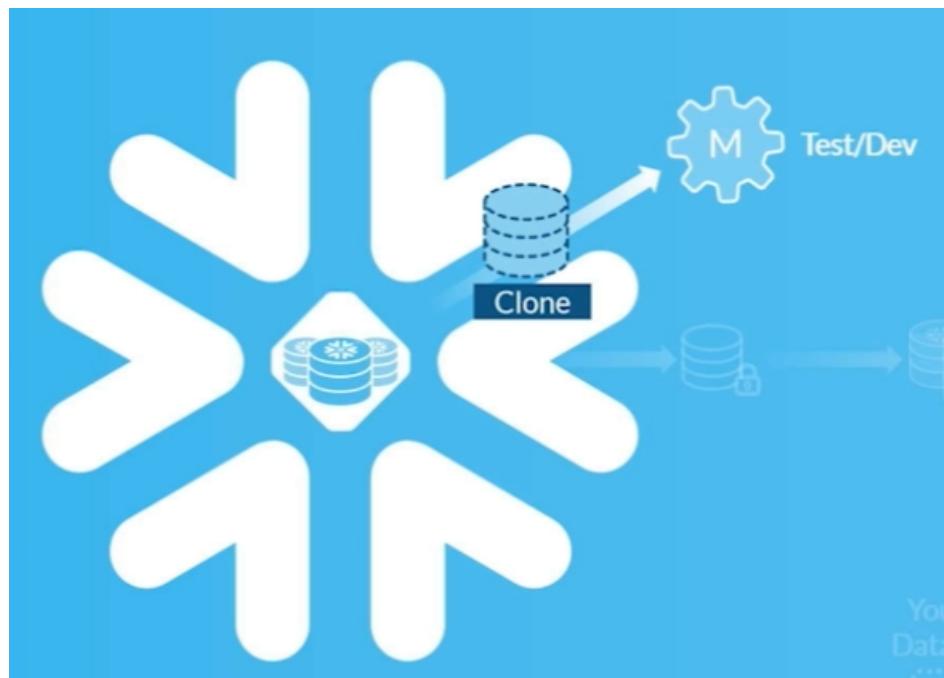


- 쿼리를 사용하는 사용자가 많거나, 실행할 쿼리가 너무 많을 경우 처리를 기다려야 할 경우, 멀티클러스터 웨어하우스, 즉 멀티 클러스터 컴퓨팅을 활용하여 해결
- 각 가상의 웨어하우스는 중앙 집중형 데이터베이스, 즉 누구나 같은 데이터에 동시 액세스가 가능

- **Snowflake의 Data sharing**



- Secure Sharing을 사용한 데이터 공유의 방식
 - Private Data를 직원과 사내 공유
 - Public Data를 교환 및 공유
 - 비즈니스 생태계에 공개
- Data Sharing 기능 예시
 - Data Marketplace 또는 Data Exchange 기능을 사용하여 데이터를 이동하거나 데이터를 복사하지 않고 공유할 수 있다.



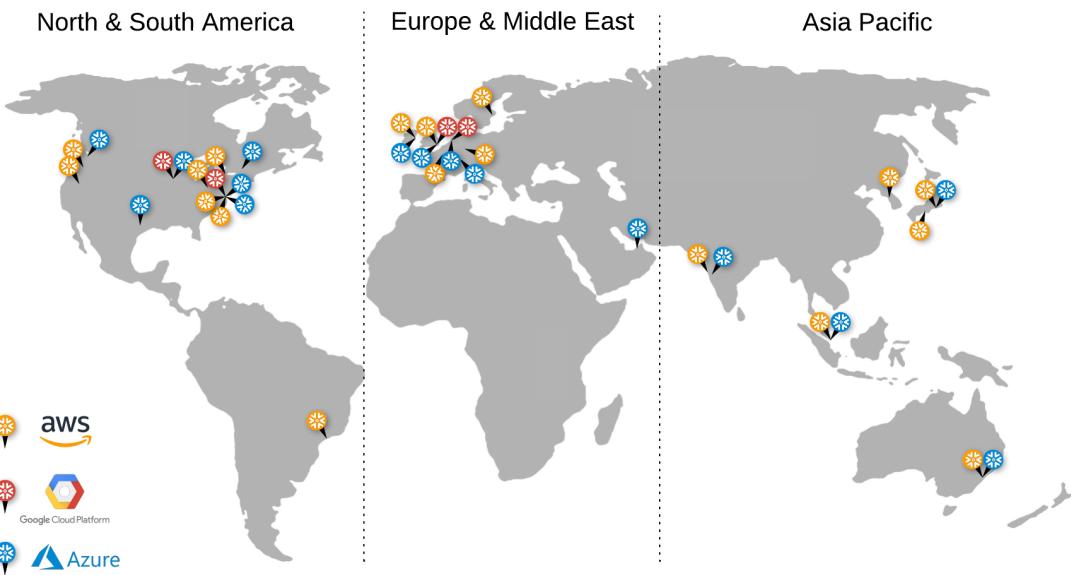
- Clone (데이터 복제, 클로닝) - 테스트 및 개발환경을 쉽고 빠르게 스핀 업에 사용

- Snowflake 의 라이선스 정책

Standard	Enterprise	Business Critical	Virtual Private Snowflake (VPS)
Complete SQL Data Warehouse Secure Data Sharing across regions / clouds Business hour support M-F 1 day of time travel Always-on enterprise grade encryption in transit and at rest Customer dedicated virtual warehouses Federated authentication Database Replication	Premier + Multi-Cluster warehouse Up to 90 days of time travel Annual rekey of all encrypted data Materialized Views AWS PrivateLink available for an extra fee	Enterprise + HIPPA support PCI compliance Data encryption everywhere Tri-Secret Secure using customer-managed keys (AWS) AWS PrivateLink support Enhanced security policy Database Failover and Fallback for business continuity	Business Critical + Customer dedicated virtual servers wherever the encryption key is in memory Customer dedicated metadata store Additional operational visibility

- Standard 부터 Business Critical 라이선스는 멀티테넌트 환경, VPS 는 단일테넌트
- 우측으로 갈수록 멀티클러스터 웨어하우스, Time Travel, PCI, HIPPA 지원 등의 기능 향상 (그림으로 확인)

- **Snowflake Supported Region**



- AWS, Azure, GCP 지원 가능 리전은 [Documentation](#) 참고
- GCP가 가장 최근에 추가된 클라우드 공급자 이기에 지원 리전이 가장 작음