

Azure Data Warehouse Architecture

에샨트 가르그

데이터 엔지니어, 아키텍트, Advisor
eshant.garg@gmail.com



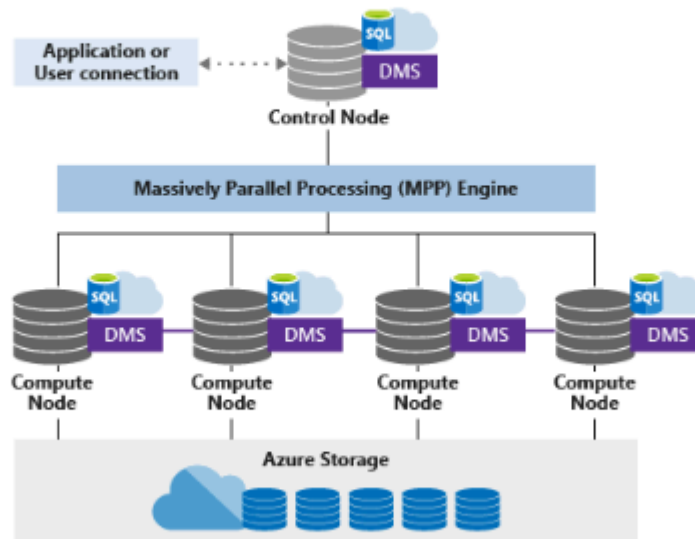
소개

MPP 또는 대규모 병렬 처리스토리지 및 데이터 배포(해시, 라운드 로빈, 복제)데이터 유형 및 테이블 유형(Columstore, Heap, Clustered B-tree 인덱스)분할 및 분산 키차원 모델링의 응용 프로그램데모 – 클라우드로 마이그레이션하기 전에 테이블 분석



Azure Synapse MPP 아키텍처

관리포트100152050034

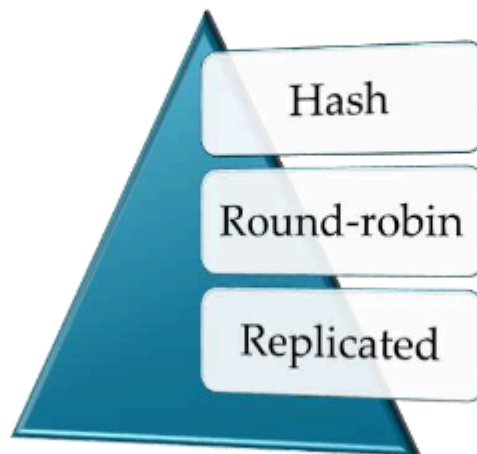


출처: Microsoft

Azure Storage 및 배포

	SQL DW는 스토리지 사용량에 대해 별도로 요금을 부과합니다
	분산은 병렬을 위한 저장 및 처리의 기본 단위입니다 쿼리
	행은 병렬로 실행되는 60개의 배포에 저장됩니다
	각 컴퓨팅 노드는 60개 배포 중 하나 이상을 관리합니다

샤딩 패턴

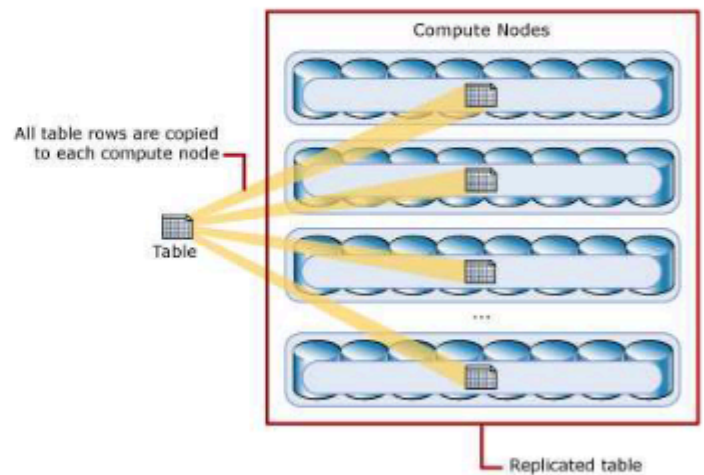


복제된 테이블

- 각 컴퓨팅 노드에서 전체 복사본을 캐시합니다.
- 작은 테이블에 사용

```
테이블 만들기 [dbo]. [비즈니스 계층](  
    [도서 ID] [nvarchar](250)  
    , [구분] [nvarchar](100) , [클러스터] [nvarchar](100) , [데스크] [nvarchar](100) , [책] [nvarchar](100) , [볼커] [nvarchar](100) , [지역] [nvarchar](100)) WITH(  
    클러스터형 COLUMNSTORE 인덱스, 분포 = REPLICATE);
```

클러스터형 COLUMNSTORE 인덱스, 분포 = REPLICATE);



라운드 로빈 테이블



테이블 만들기 [dbo]. [일정](
[일시] [날짜/시간2](3), [DateKey] [10진수](38,
0), [요일] [nvarchar](100), [일] [소수점](38,
0))

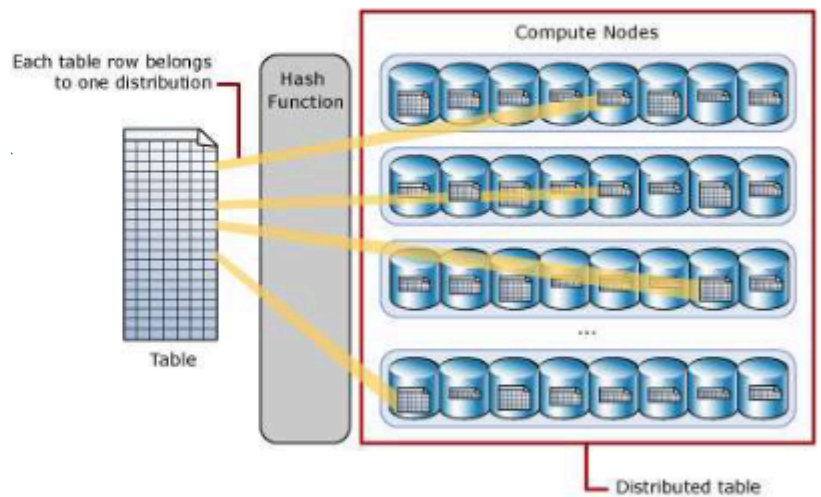
위드(
클러스터형 COLUMNSTORE 인덱스, 분
포 = ROUND_ROBIN);

- 일반적으로 스테이징 테이블을 로드하는 데 사용합니다.
- 테이블 전체에 걸쳐 데이터를 균등하게 분배
추가 최적화
- 조인은 데이터를 다시 섞어야 하기 때문에 속도가 느립니다.
- 기본 배포 유형

출처: Microsoft

해시 분산 테이블

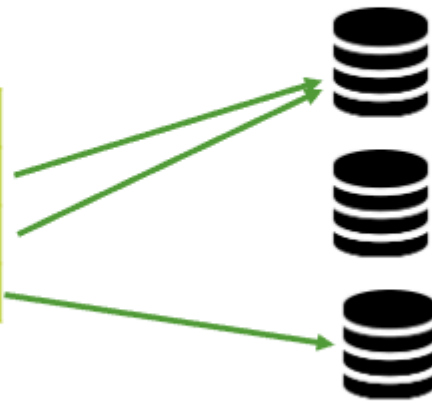
- 대형 테이블을 위한 최고의 성능
- 각 행이 하나의 특정 분포에 속합니다.
- 주로 큰 테이블에 사용됩니다.



출처: Microsoft

해시 분산 테이블

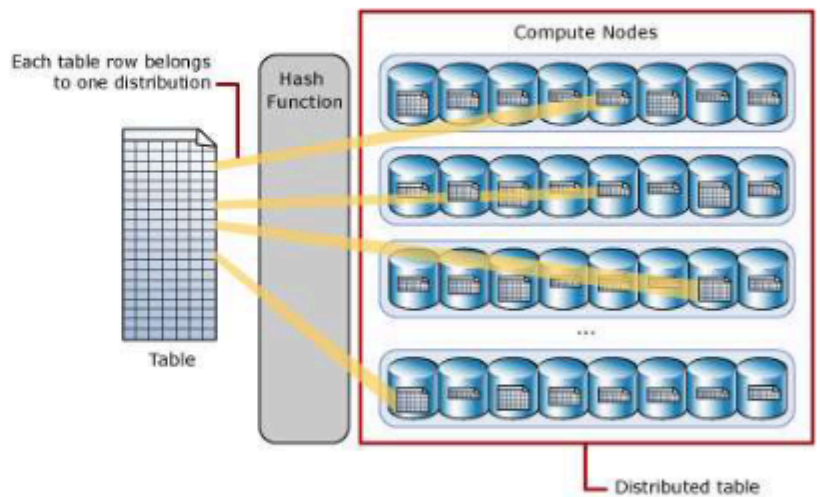
RecordProductStore1 축구뉴욕		
2	축구	로스앤젤레스
3	축구피닉스	



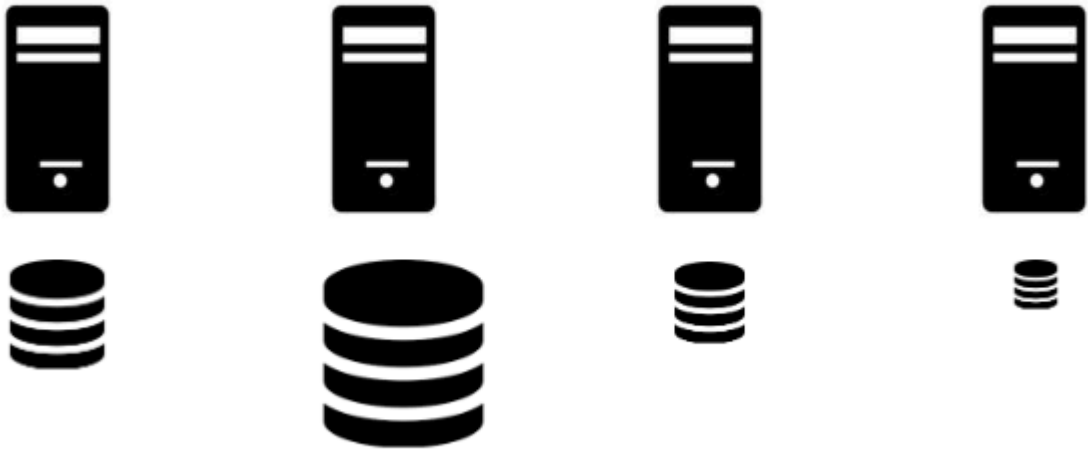
해시 분산 테이블

- 대형 테이블을 위한 최고의 성능
- 각 행은 하나의 특정 행에 속합니다.
- 분포
- 주로 큰 테이블에 사용됩니다.

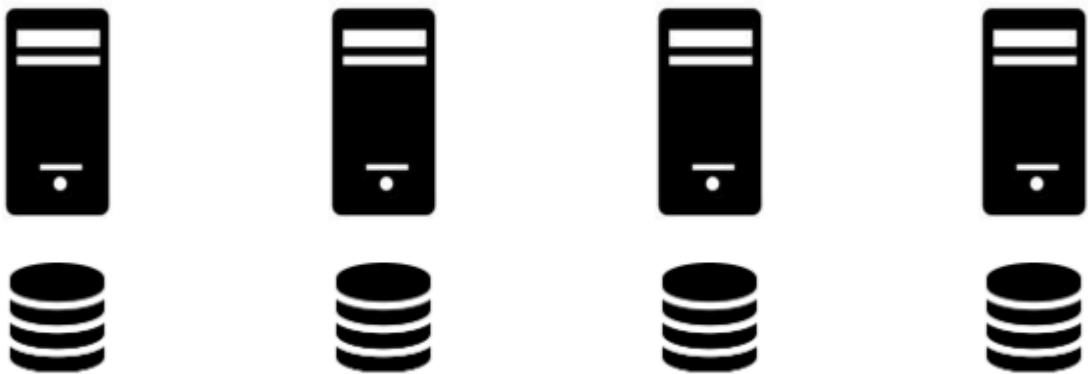
테이블 만들기 [dbo].[주식TimeSeries데이 터]([날짜] [varchar](30),[BookId] [10진수] (38, 0),[P&L] [10진수](31, 7),[VaRLower] [10진수](31, 7))WITH(클러스터형 COLUMNSTORE 인덱스, 분포 = HASH([P&L]));



데이터 스쿠 방지



균등 분포

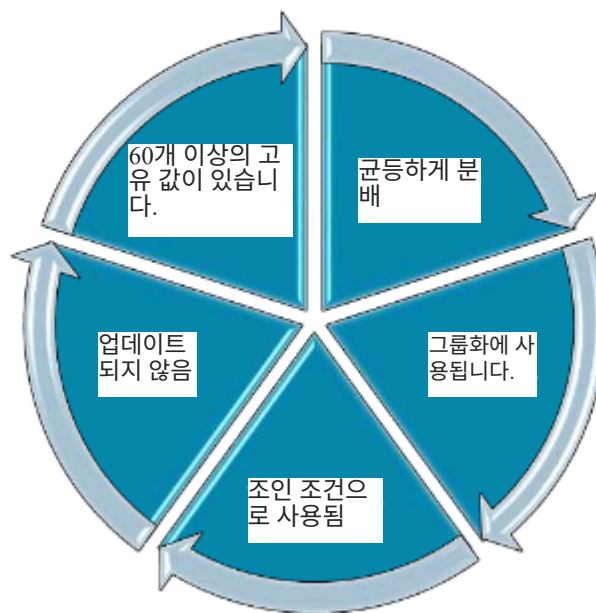


Distribution key

Azure SQL Data Warehouse가 데이터를 분산하는 방법을 결정합니다.
여러 노드에 걸쳐.

Azure SQL Data Warehouse는 데이터를 로드할 때 최대 60개의 배포를 사용합니다.
체계.


양호한 해시 키



어떤 데이터 배포를 사용할 것인가?

형	잘 맞는곳조심하세요...	
복제	astar 스키마의 작은 차원 테이블과 2GB 미만의 스토리지 aftercompression	<ul style="list-style-type: none"> • 많은 쓰기 트랜잭션이 테이블에 있습니다. (삽입/업데이트/삭제) • DWU 프로비저닝을 자주 변경하는 경우 • 2-3개의 열만 사용하지만 테이블에는 많은 열 • 복제된 테이블을 인덱싱합니다.
라운드 로빈(기본값)	<ul style="list-style-type: none"> • 임시/스테이징 테이블 • 명백한 결합 키가 없거나 좋은 후보 칼럼. 	데이터 이동으로 인해 성능이 느려집니다.
해시	<ul style="list-style-type: none"> • 팩트 테이블 • 큰 차원 테이블 	배포 키는 업데이트할 수 없습니다

데이터 유형

	데이터를 지원하는 가장 작은 데이터 유형 사용
	모든 문자 열을 큰 기본값으로 정의하지 마십시오. 길이
	다음과 같은 경우 열을 NVARCHAR가 아닌 VARCHAR로 정의합니다. 유니코드가 필요하지 않습니다.

데이터 유형



목표는 공간을 절약할 뿐만 아니라 데이터를 최대한 효율적으로 이동하는 것입니다.

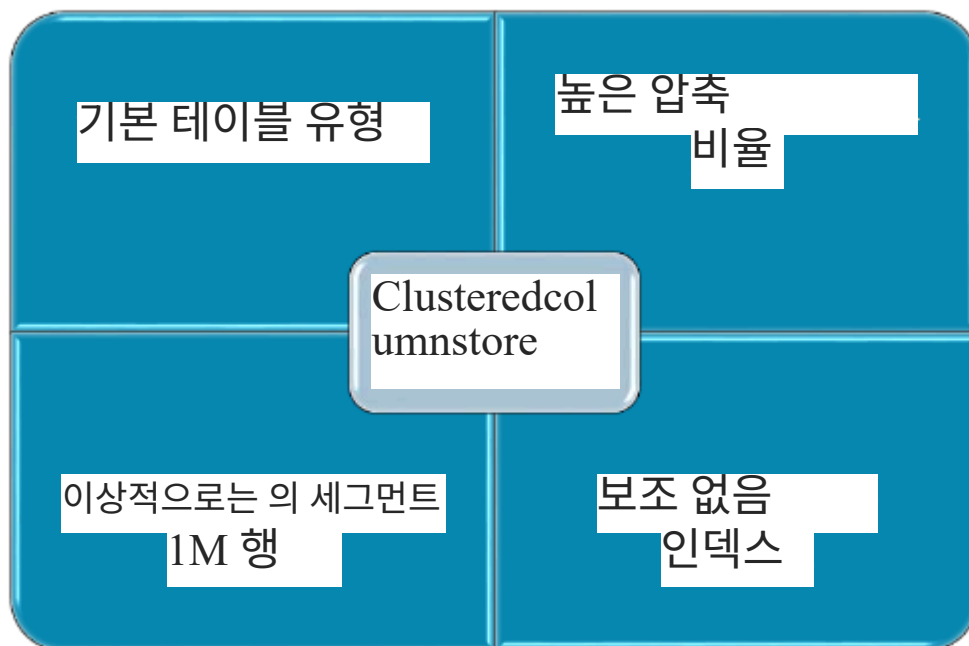
데이터 유형

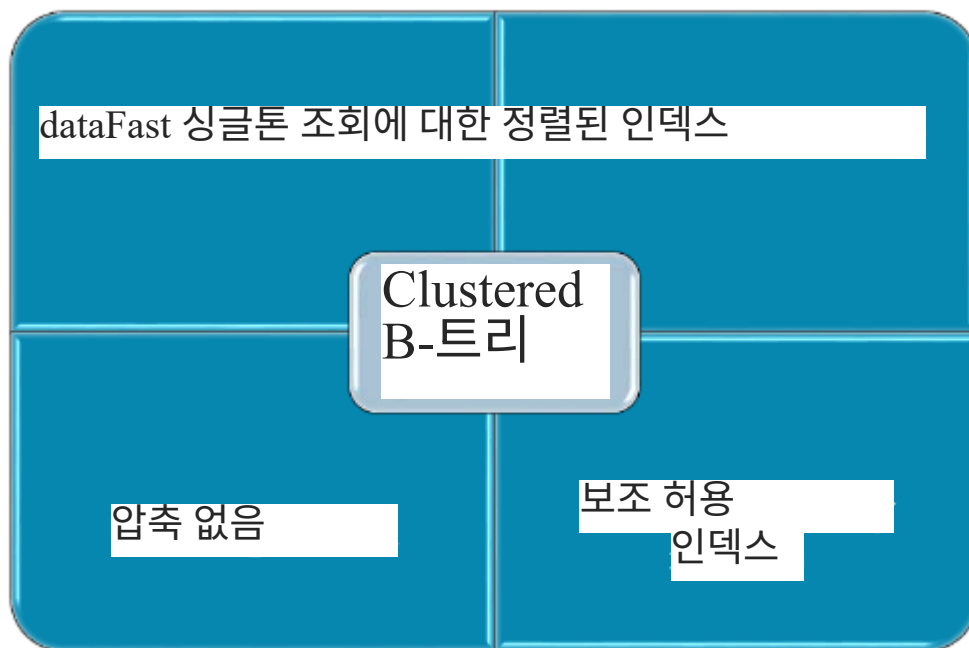
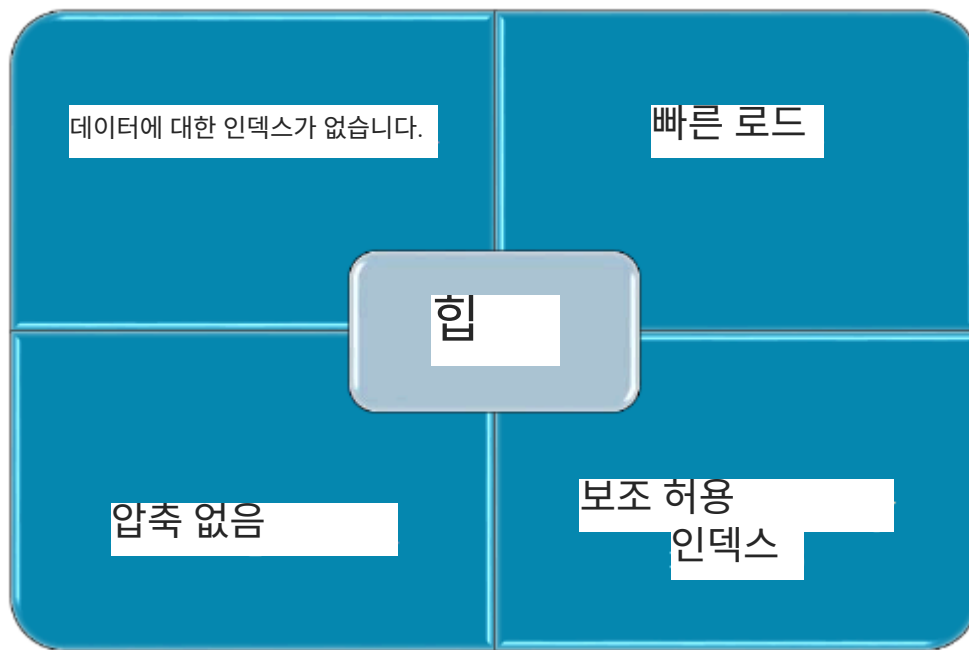


일부 복잡한 데이터 유형(XML, 지리 등)
Azure SQL 데이터에서 지원되지 않습니다.
아직 참고가 있습니다.

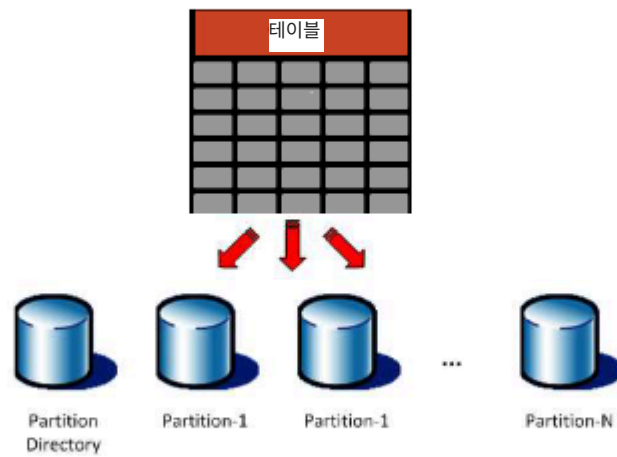
테이블 유형

Clusteredcolumnstore	<ul style="list-style-type: none">업데이트 가능한 기본 저장 방법읽기 전용에 적합
힙	<ul style="list-style-type: none">데이터는 특정 순서가 없습니다.데이터에 자연스러운 순서가 없는 경우 사용합니다.
클러스터형 인덱스	<ul style="list-style-type: none">인덱싱되는 데이터와 물리적으로 동일한 순서로 저장된 인덱스입니다





테이블 파티셔닝



분할

테이블 파티션을 사용하면 데이터를 다음과 같이 나눌 수 있습니다.
더 작은 데이터 그룹

데이터 로드의 효율성과 성능 향상 파티션 삭제, 전환 및 병합을 사용하여 일반적으로 데이터는 언제 연결된 날짜 열에서 분할됩니다.

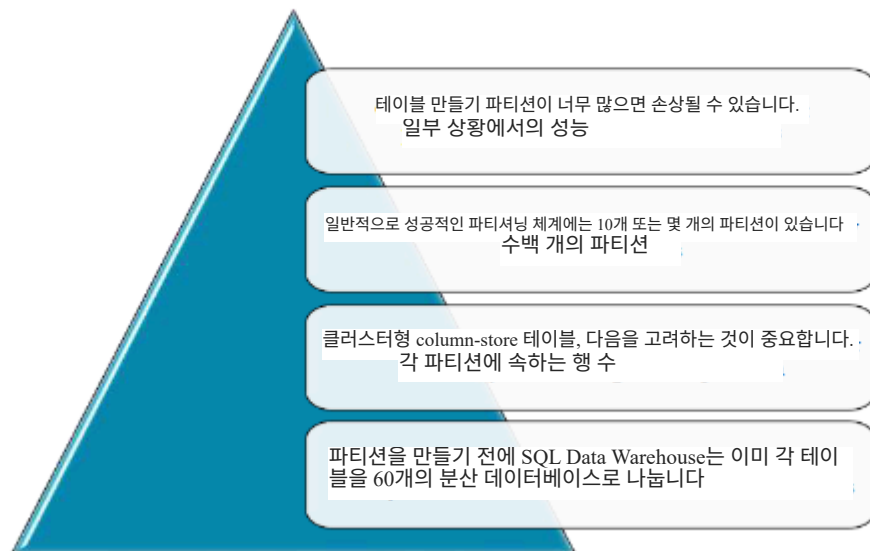
데이터가 데이터베이스에 로드됩니다.

쿼리 성능을 개선하는 데에도 사용할 수 있습니다.

파티셔닝이 필요한 이유



파티션에 대한 유용한 정보





매우 세분화된 분할 체계는 SQL Server에서 작동할 수 있지만 Azure에서는 성능을 저하시킬 수 있습니다
SQL 데이터 웨어하우스.

본보기

60 배포판365 파티션



21900 데이터 버킷

21900 데이터 버킷



이상적인 SegmentSize(1M 행)




21 900 000 000 행




낮은 세분성(주, 월)은 보유한 데이터의 양에 따라 더 나은 성능을 발휘할 수 있습니다.

How do we apply these principles to a Dimensional model?

팩트 테이블

	큰 것은 Columnstore로 더 좋습니다.
	해시 키를 통해 배포되는 만큼 테이블이 각 세그먼트를 채울 수 있을 만큼 충분히 큰 경우에만 Partitioned인 한 가능합니다

차원 테이블

	해시 분산 또는 명확한 후보 조인 키가 없는 경우 라운드 로빈일 수 있습니다.
	큰 차원을 위한 Columnstore
	작은 차원을 위한 힙 또는 클러스터형 인덱스Heap or Clustered Index for small dimensions
	대체 조인 열에 대한 보조 인덱스 추가Add secondary indexes for alternate join columns
	파티셔닝은 권장되지 않습니다.

데모

Azure Synapse 데이터 풀로 마이그레이션하기 전에 온-프레미스 Datawarehouse에서 데이터 배포를 분석합니다.

- Microsoft의 AdventureworksDW 데이터베이스를 온-프레미스 데이터 웨어하우스로 사용합니다.
- 우리는 하나의 차원과 하나의 팩트 테이블을 분석할 것입니다.
- 온-프레미스 데이터베이스의 다른 테이블에도 동일한 프로세스를 반복할 수 있습니다.

요약

MPP 또는 대규모 병렬 처리청구 = 컴퓨팅 + 스토리지
데이터 배포(해시, 라운드 로빈, 복제)데이터 유형 및
테이블 유형데이터 분할모범 사례 – 팩트 및 차원 테
이블 설계데모 – 데이터 분포 분석



Azure
Synapse
Analytics