

SQL을 이용한 데이터 분석

1. SQL과 데이터베이스 소개

한기용

keeyonghan@hotmail.com

Contents

1. SQL의 중요성
2. 강사 소개
3. 배움이란?
4. 관계형 데이터베이스란?
5. SQL이란?
6. 데이터 웨어하우스란?
7. Cloud와 AWS 소개
8. Redshift 소개

SQL의 중요성

모든 데이터 직군에게 필요한 기술은 SQL

◆ 데이터 관련 3개의 직군

❖ 데이터 엔지니어:

- 파이썬, 자바/스칼라
- **SQL**, 데이터베이스
- ETL/ELT (Airflow, DBT)
- Spark, Hadoop

❖ 데이터 분석가

- **SQL**, 비즈니스 도메인에 대한 지식
- 통계 (AB 테스트 분석)

❖ 데이터 과학자

- 머신러닝
- **SQL**, 파이썬
- 통계

데이터 요약과 데이터 분석을 위한
SQL

◆ 강사 소개 - 한기용

- ❖ 2020-현재 Harmonize Health 데이터 팀 디렉터
- ❖ 2018-2020년 데이터 관련 컨설턴트/어드바이저/엔젤투자자
- ❖ 2014-2018년 Udemy 데이터 팀 시니어 디렉터
- ❖ 2012-2014년 Polyvore 데이터 팀 시니어 매니저
- ❖ 2004-2011년 Yahoo Search 엔지니어링 디렉터
- ❖ 2000년에 미국 실리콘밸리로 이주
- ❖ 1995-2000년 삼성전자 소프트웨어 개발자
- ❖ 서울대학교 컴퓨터 공학과 학사/석사

배움이란?

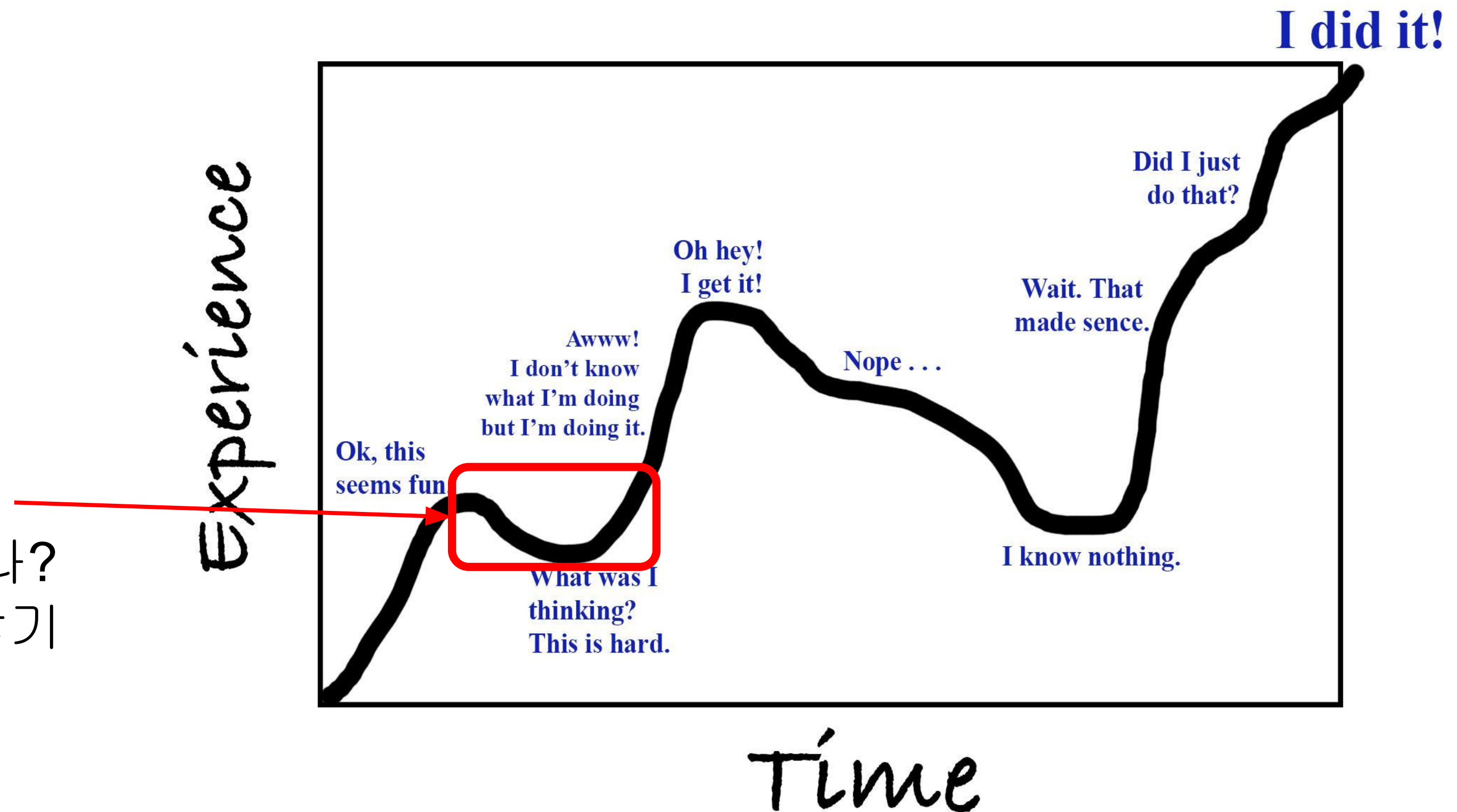
배움에는 시간과 노력이 걸림

◆ 배움의 전형적인 패턴

The Learning Curve

여기서 어떻게 하느냐가 아주 중요!

1. 가장 중요한 것은 버티는 힘
 - a. 이걸 즐겨야함 :)
2. 내가 뭘 모르지는지 생각해봐야함
 - a. 내가 어디서 막혔는지 구체적으로 질문할 수 있나?
3. 잘 하는 사람 보고 기죽지 않기



◆ 새로운 것을 처음 배울 때의 좋은 자세 (1)

❖ 자신이 아는 것과 모르는 것을 분명히 이해하는지?

- 멍청한 질문이란 없다는 점 명심. 대충 알거나 모르면서 안 물어보는 것이 더 큰 문제
- 이는 피드백을 잘 받아들일 수 있는지와도 연계됨

❖ 마음을 편하게 먹기

- 내가 이해하기 힘들다면 남들도 이해하기 힘들
- 나보다 잘 하는 사람들은 그만큼 시간을 쏟았기 때문

◆ 새로운 것을 처음 배울 때의 좋은 자세 (2)

❖ 배움의 발전은 tipping point를 거치면서 폭발하는 형태임

- "The secret is to build the resolve and spirit to enjoy the plateaus the times when it doesn't feel like you're improving and you question why you are doing this. If you're patient, the plateaus will become springboards" (Quotes from Steve Nash)
- 발전이 더딘 기간을 **즐기는** 자세가 필요

관계형 데이터베이스란?

구조화된 데이터를 저장하는데 사용되는 관계형
데이터베이스가 무엇인지 알아보자

◆ 관계형 데이터베이스

- ❖ 구조화된 데이터를 저장하고 질의할 수 있도록 해주는 스토리지
 - 엑셀 스프레드시트 형태의 테이블로 데이터를 정의하고 저장
 - 테이블에는 컬럼(열)과 레코드(행)이 존재
- ❖ 관계형 데이터베이스를 조작하는 프로그래밍 언어가 **SQL**
 - 테이블 정의를 위한 **DDL (Data Definition Language)**
 - 테이블 데이터 조작/질의를 위한 **DML (Data Manipulation Language)**

◆ 대표적 관계형 데이터베이스

❖ 프로덕션 데이터베이스: MySQL, PostgreSQL, Oracle, ...

- OLTP (OnLine Transaction Processing)
- 빠른 속도에 집중. 서비스에 필요한 정보 저장

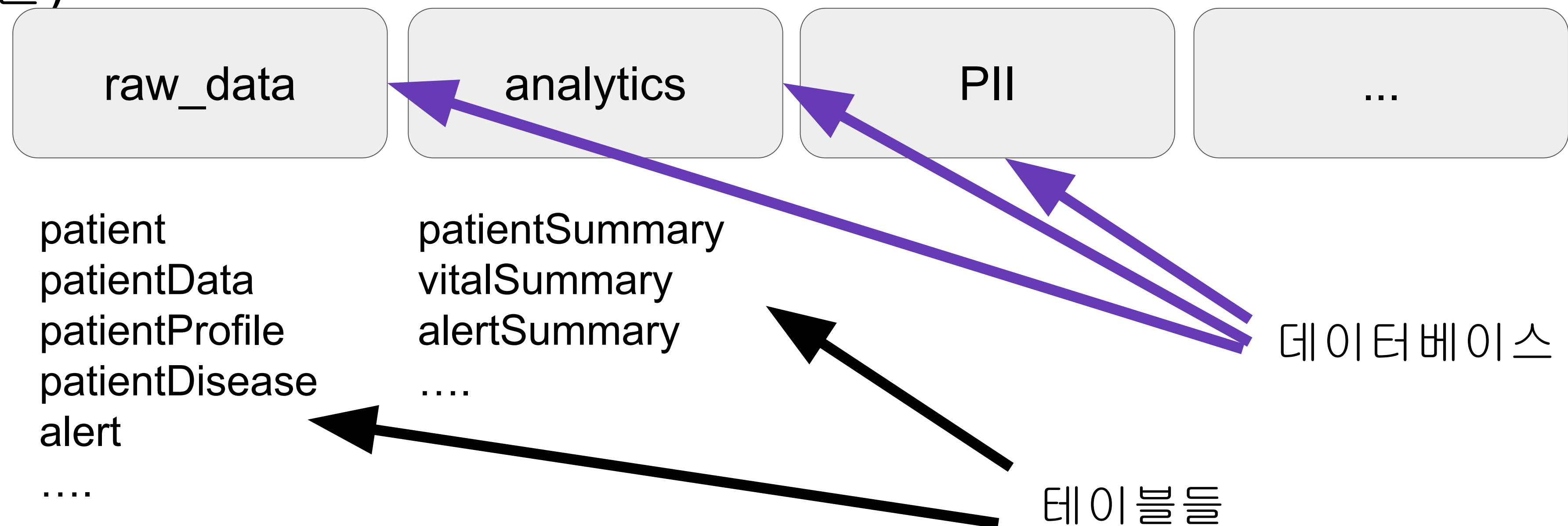
❖ 데이터 웨어하우스: Redshift, Snowflake, BigQuery, Hive, ...

- OLAP (OnLine Analytical Processing)
- 처리 데이터 크기에 집중. 데이터 분석 혹은 모델 빌딩등을 위한 데이터 저장
 - 보통 프로덕션 데이터베이스를 복사해서 데이터 웨어하우스에 저장

◆ 관계형 데이터베이스의 구조

❖ 관계형 데이터베이스는 2 단계로 구성됨

- 가장 밑단에는 테이블들이 존재 (테이블은 엑셀의 시트에 해당)
- 테이블들은 데이터베이스(혹은 스키마)라는 폴더 밑으로 구성 (엑셀에서는 파일)



◆ 관계형 데이터베이스의 구조

❖ 테이블의 구조 (테이블 스키마라고 부르기도 함)

- 테이블은 레코드들로 구성 (행)
- 레코드는 하나 이상의 필드(컬럼)로 구성 (열)
- 필드(컬럼)는 이름과 타입과 속성(primary key)으로 구성됨

컬럼	타입
userId	int
sessionId	varchar(32)
channel	varchar(32)

테이블 스키마

userId	sessionId	channel
779	7cdace91c487558e27ce54df7cdb299c	Instagram
230	94f192dee566b018e0acf31e1f99a2d9	Naver
369	7ed2d3454c5eea71148b11d0c25104ff	Youtube
248	f1daf122cde863010844459363cd31db	Naver

테이블 레코드 예

SQL이란?

데이터 처리에서 기본이 되는 SQL에 대해 배우자!

◆ SQL 소개

❖ SQL: Structured Query Language

- 관계형 데이터베이스에 있는 데이터(테이블)를 질의하거나 조작해주는 언어

❖ SQL은 1970년대 초반에 IBM이 개발한 구조화된 데이터 질의 언어

❖ 두 종류의 언어로 구성됨

- DDL (Data Definition Language):

- 테이블의 구조를 정의하는 언어

- DML (Data Manipulation Language):

- 테이블에서 원하는 레코드들을 읽어오는 질의 언어
- 테이블에 레코드를 추가/삭제/갱신해 주는데 사용하는 언어

◆ SQL은 빅데이터 세상에서도 중요!

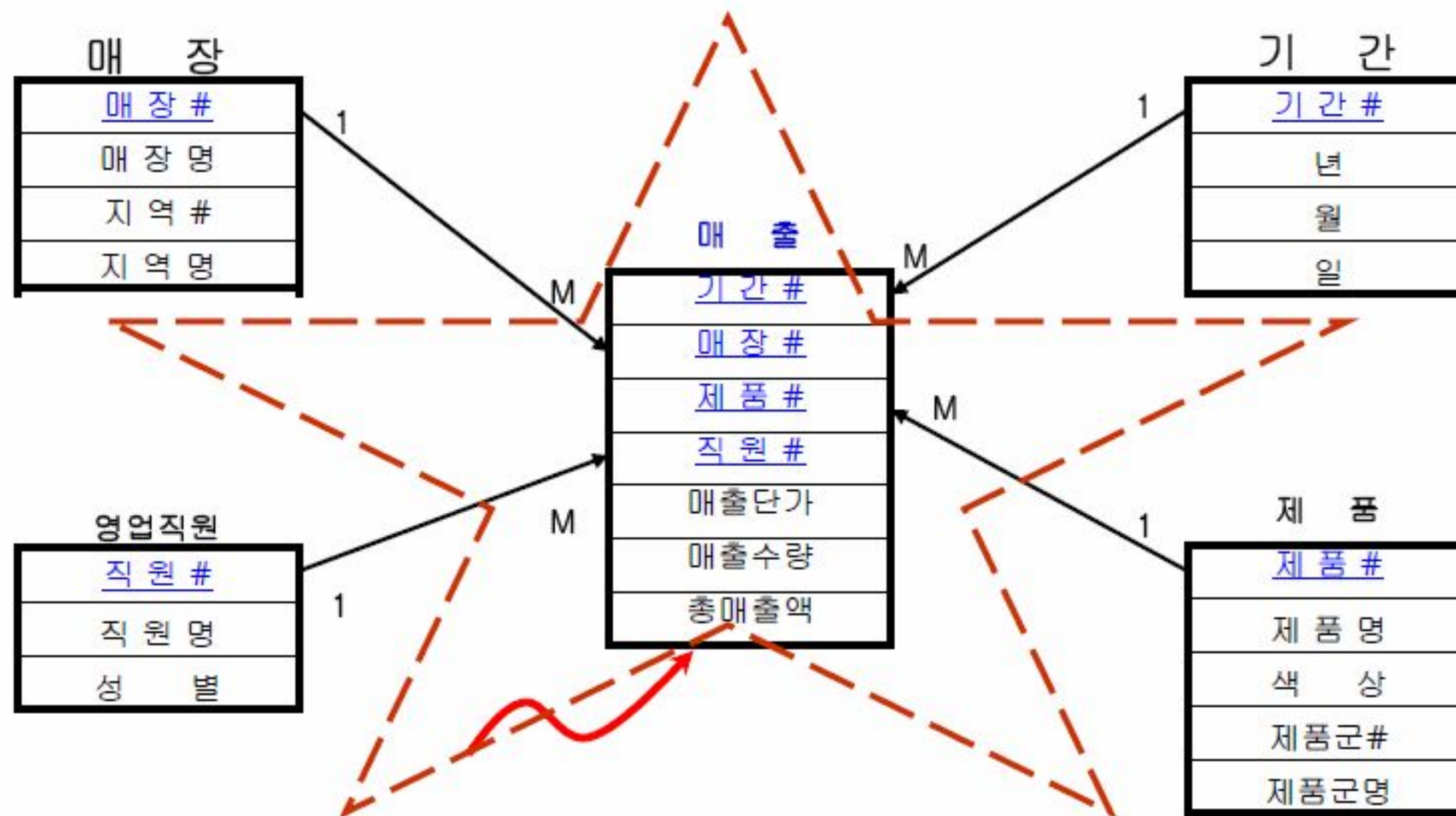
- ❖ 구조화된 데이터를 다루는 SQL은 데이터 규모와 상관없이 쓰임
- ❖ 모든 대용량 데이터 웨어하우스는 SQL 기반
 - Redshift, Snowflake, BigQuery, Hive
- ❖ Spark이나 Hadoop도 예외는 아님
 - SparkSQL과 Hive라는 SQL 언어가 지원됨
- ❖ 데이터 분야에서 일하고자 하면 반드시 익혀야할 기본 기술
 - 데이터 엔지니어, 데이터 분석가, 데이터 과학자 모두 알아야함

◆ SQL의 단점

- ❖ 구조화된 데이터를 다루는데 최적화가 되어있음
 - 정규표현식을 통해 비구조화된 데이터를 어느 정도 다루는 것은 가능하나 제약이 심함
 - 많은 관계형 데이터베이스들이 플랫폼 구조만 지원함 (no nested like JSON)
 - 구글 빅쿼리는 **nested structure**를 지원함
 - 비구조화된 데이터를 다루는데 **Spark, Hadoop**과 같은 분산 컴퓨팅 환경이 필요해짐
 - 즉 **SQL**만으로는 비구조화 데이터를 처리하지 못함
- ❖ 관계형 데이터베이스마다 **SQL** 문법이 조금씩 상이

◆ Star schema

- ❖ Production DB용 관계형 데이터베이스에서는 보통 스타 스키마를 사용해 데이터를 저장
- ❖ 데이터를 논리적 단위로 나눠 저장하고 필요시 조인. 스토리지의 낭비가 덜하고 업데이트가 쉬움



◆ Denormalized schema

- ❖ 데이터 웨어하우스에서 사용하는 방식
 - 단위 테이블로 나눠 저장하지 않음으로 별도의 조인이 필요 없는 형태를 말함
- ❖ 이는 스토리지를 더 사용하지만 조인이 필요 없기에 빠른 계산이 가능

년 월 일
매장명
지역명
직원명
성별
매출단가
매출수량
총매출액
제품명
색상
제품군명

Denormalize된
매출 테이블

데이터 웨어하우스 소개

데이터 웨어하우스가 무엇이고 다른 관계형 데이터베이스와
어떻게 다른지 알아보자

◆ 데이터 웨어하우스: 회사에 필요한 모든 데이터를 저장

❖ 여전히 SQL 기반의 관계형 데이터베이스

- 프로덕션 데이터베이스와는 별도로이어야 함
 - OLAP (OnLine Analytical Processing) vs. OLTP (OnLine Transaction Processing)
- AWS의 Redshift, Google Cloud의 Big Query, Snowflake 등이 대표적
 - 고정비용 옵션 vs. 가변비용 옵션

❖ 데이터 웨어하우스는 고객이 아닌 내부 직원을 위한 데이터베이스

- 처리속도가 아닌 처리 데이터의 크기가 더 중요해짐

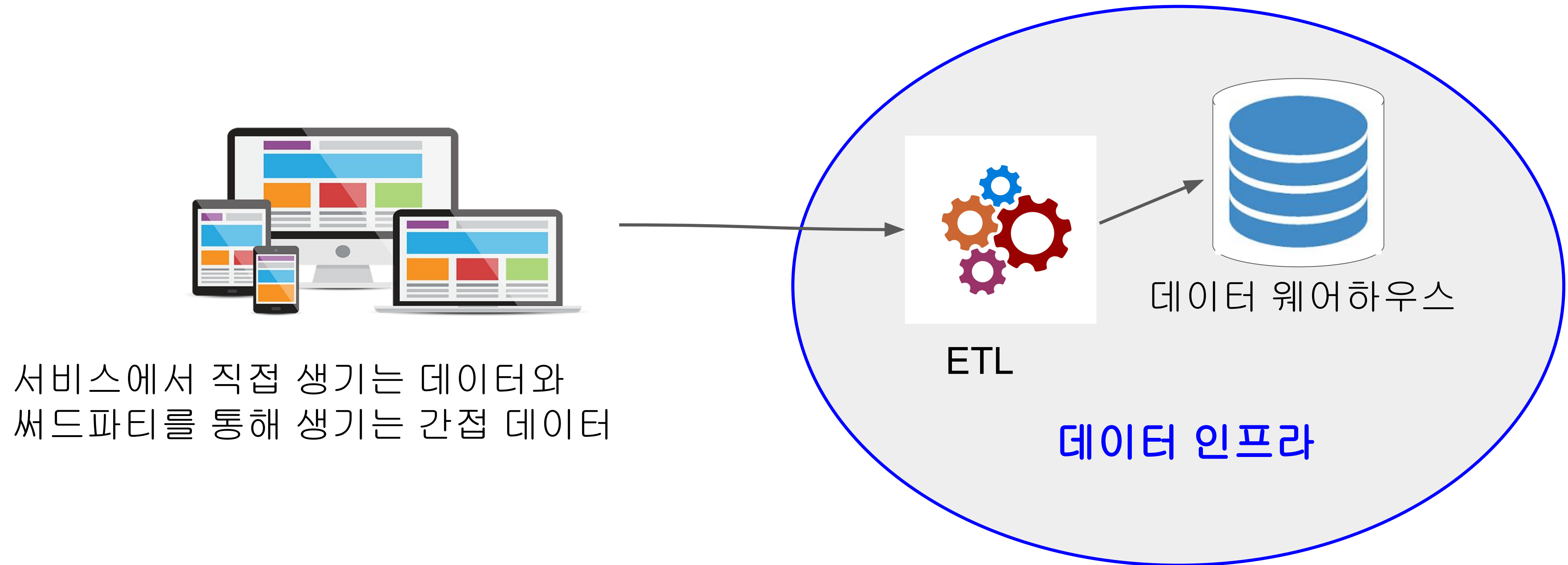
❖ ETL 혹은 데이터 파이프라인

- 외부에 존재하는 데이터를 읽어다가 데이터 웨어하우스로 저장해주는 코드들이 필요해지는데 이를 ETL 혹은 데이터 파이프라인이라고 부름

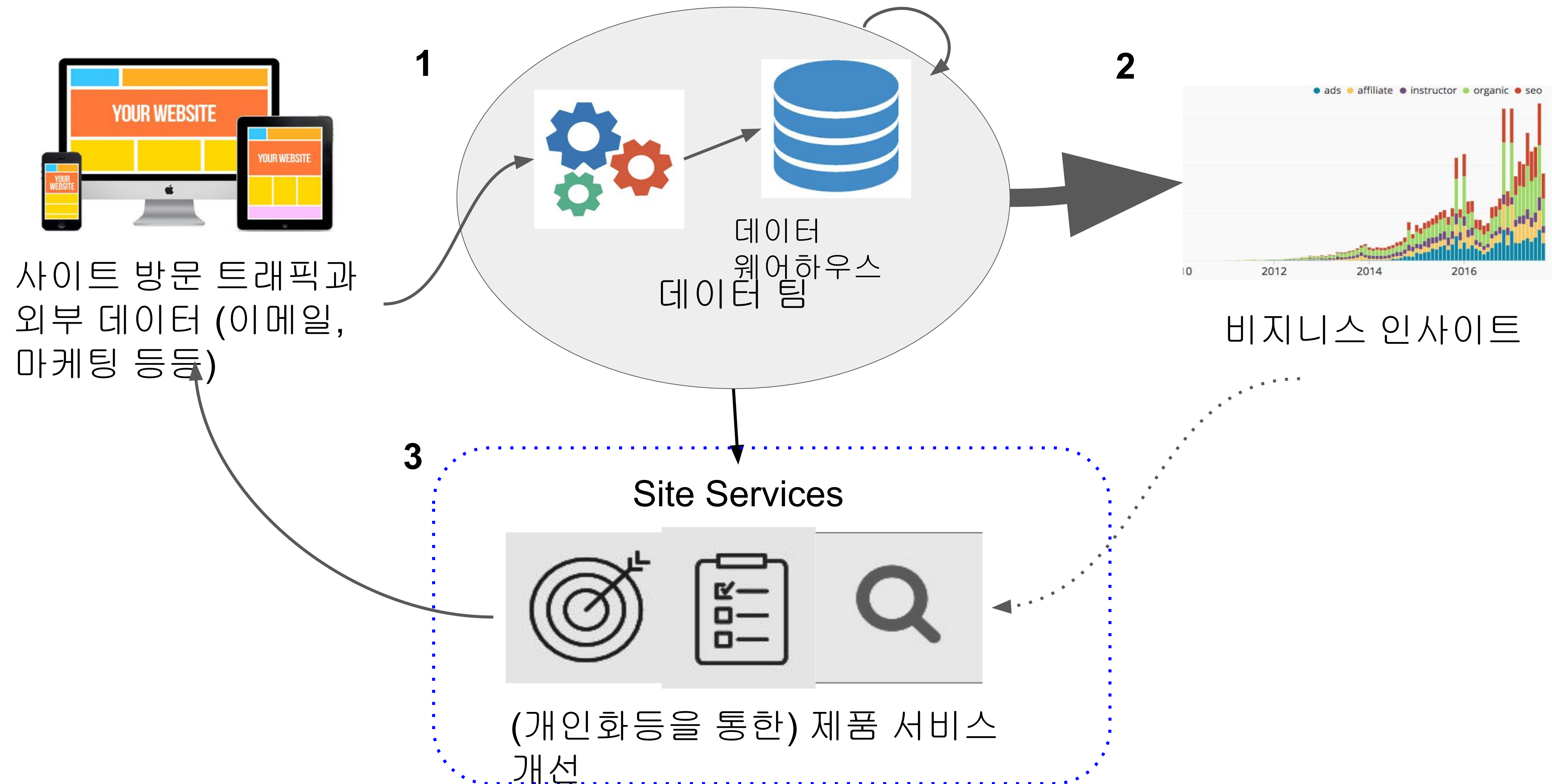
◆ 데이터 인프라란?

❖ 데이터 엔지니어가 관리함

- 여기서 한 단계 더 발전하면 Spark과 같은 대용량 분산처리 시스템이 일부로 추가됨



◆ 데이터 순환 구조



Cloud와 AWS 소개

Cloud와 AWS에 대해 간략히 배워보자

◆ 클라우드의 정의

- ❖ 컴퓨팅 자원(하드웨어, 소프트웨어 등등)을 네트워크를 통해 서비스 형태로 사용하는 것.
- ❖ 키워드:
 - “No Provisioning”
 - “Pay As You Go”
- ❖ 자원(예를 들면 서버)을 필요한만큼 (거의) 실시간으로 할당하여 사용한만큼 지불
 - 탄력적으로 필요한만큼의 자원을 유지하는 것이 중요

◆ 클라우드 컴퓨팅이 없었다면?

- ❖ 서버/네트워크/스토리지 구매와 설정등을 직접 수행해야 함
- ❖ 데이터센터 공간을 직접 확보 (Co-location)
 - 확장이 필요한 경우 공간을 먼저 더 확보해야함
- ❖ 그 공간에 서버를 구매하여 설치하고 네트워크 설정
 - 보통 서버를 구매해서 설치하는데 적어도 두세달은 걸림
- ❖ 또한 **Peak time**을 기준으로 **Capacity planning**을 해야함!
 - 놓고 있는 자원들이 높게 되는 현상 발생
- ❖ 직접 운영비용 **vs.** 클라우드 비용
 - 기회비용!

◆ 클라우드 컴퓨팅의 장점

- ❖ 초기 투자 비용이 크게 줄어듦
 - CAPEX (Capital Expenditure) vs. OPEX (Operating Expense)
- ❖ 리소스 준비를 위한 대기시간 대폭 감소
 - Shorter Time to Market
- ❖ 노는 리소스 제거로 비용 감소
- ❖ 글로벌 확장 용이
- ❖ 소프트웨어 개발 시간 단축
 - Managed Service (SaaS) 이용

◆ AWS 소개

- ❖ 가장 큰 클라우드 컴퓨팅 서비스 업체.
- ❖ 2002년 아마존의 상품데이터를 **API**로 제공하면서 시작
 - 현재 100여개의 서비스를 전세계 15개의 지역에서 제공.
 - 대부분의 서비스들이 오픈소스 프로젝트들을 기반으로 함.
 - 최근 들어 **ML/AI** 관련 서비스들도 내놓기 시작
- ❖ 사용고객
 - Netflix, Zynga등의 상장업체들도 사용.
 - 많은 국내 업체들도 사용시작 (서울 리전)
- ❖ 다양한 종류의 소프트웨어/플랫폼 서비스를 제공.
 - **AWS**의 서비스만으로 쉽게 온라인서비스 생성.
 - 뒤에서 일부 서비스를 따로 설명.

Amazon Web Services, 5-year financials



Source: Amazon Earnings Reports, In Millions Per Fiscal Quarter

GEEKWIRE


AWS Regions

Regions	Name
US East (Virginia)	us-east-1
US East (Ohio)	us-east-2
US West (N. California)	us-west-1
US West (Oregon)	us-west-2
Asia Pacific (Mumbai)	ap-south-1
Asia Pacific (Osaka)	ap-northeast-3
Asia Pacific (Seoul)	ap-northwest-2
Asia Pacific (Singapore)	ap-southeast-1
Asia Pacific (Sydney)	ap-southeast-2
Asia Pacific (Tokyo)	ap-northeast-1
Canada (Central)	ca-central-1
China (Beijing)	cn-north-1
China (Ningxia)	cn-northwest-1
EU (Frankfurt)	eu-central-1
EU (Ireland)	eu-west-1
EU (Paris)	eu-west-3
EU (Stockholm)	eu-north-1
EU (London)	eu-west-2
South America (Sao Paulo)	sa-east-1
AWS GovCloud (US, US-East)	us-gov-west-1, us-gov-east-1

AWS 서비스들


Search services

Group

 **Compute**

EC2


EC2 Container Service

Lightsail 

Elastic Beanstalk

Lambda

Batch


 **Storage**

S3

EFS

Glacier

Storage Gateway


 **Database**

RDS

DynamoDB

ElastiCache

Redshift


 **Networking & Content Delivery**

VPC

CloudFront

Direct Connect

Route 53


 **Developer Tools**

CodeCommit

CodeBuild

CodeDeploy

CodePipeline

 **Management Tools**

CloudWatch

CloudFormation

CloudTrail

Config


OpsWorks

Service Catalog

Trusted Advisor

Managed Services

Application Discovery Service

 **Security, Identity & Compliance**

IAM


Inspector

Certificate Manager

Directory Service

WAF & Shield

Compliance Reports

 **Analytics**

Athena


EMR


CloudSearch

Elasticsearch Service

Kinesis

Data Pipeline

QuickSight 


 **Artificial Intelligence**

Lex


Polly

Rekognition


Machine Learning

 **Internet Of Things**

AWS IoT

 **Game Development**


GameLift

 **Mobile Services**

Mobile Hub

Cognito

Device Farm


 **Application Services**

Step Functions

SWF

API Gateway


Elastic Transcoder

 **Messaging**

SQS


SNS


SES

 **Business Productivity**

WorkDocs

WorkMail

Amazon Chime 

 **Desktop & App Streaming**

WorkSpaces

AppStream 2.0

EC2 – Elastic Compute Cloud (1)

- AWS의 서버 호스팅 서비스.
 - 리눅스 혹은 윈도우 서버를 론치하고 어카운트를 생성하여 로그인 가능 (구글앱엔진과의 가장 큰 차이점).
 - 가상 서버들이라 전용서버에 비해 성능이 떨어짐.
 - Bare-metal 서버도 제공하기 시작
- 다양한 종류의 서버 타입 제공
 - <http://aws.amazon.com/ec2/>
 - 예를 들어 미국 동부에서 스몰타입(t2.small)의 무료 리눅스 서버를 하나 할당시
 - 시간당 2.3 센트의 비용지불.
 - 2GB 메모리, 1 가상코어, 160GB 하드디스크
 - 2012년에는 8.5 센트였음
 - 타입별 지역별 가격을 알고 싶다면 여기를 방문
 - Incoming network bandwidth는 공짜이지만 outgoing은 유료.

EC2 – Elastic Compute Cloud (2)

- 세 가지 종류의 구매 옵션

구매 옵션	설명
On-Demand	시간당 비용을 지불되며 가장 흔히 사용하는 옵션
Reserved	1년이나 3년간 사용을 보장하고 1/3 정도에서 40% 디스카운트를 받는 옵션
Spot Instance	일종의 경매방식으로 놓고 있는 리소스들을 보다 싼 비용으로 사용할 수 있는 옵션

S3 – Simple Storage Service (1)

- <http://aws.amazon.com/s3/>
- 아마존이 제공하는 대용량 클라우드 스토리지 서비스
- S3는 데이터 저장관리를 위해 계층적 구조를 제공
- 글로벌 네임스페이스를 제공하기 때문에 톱레벨 디렉토리 이름 선정에 주의.
- S3에서는 디렉토리를 버킷(**Bucket**)이라고 부름
- 버킷이나 파일별로 액세스 컨트롤 가능

S3 – Simple Storage Service (2)

- <https://aws.amazon.com/ko/s3/pricing/>
- Low cost. 1TB per month:
 - Standard storage: \$23
 - Infrequent Access storage: \$12.5
 - SLA가 다름
 - Glacier storage: \$4

기타 중요 서비스 - Database Services

- RDS (Relational Database Service)
 - MySQL, PostgreSQL, Aurora
 - Oracle, MS SQL Server
- DynamoDB
- **Redshift**
- ElastiCache
- Neptune (Graph database)
- ElasticSearch
- MongoDB

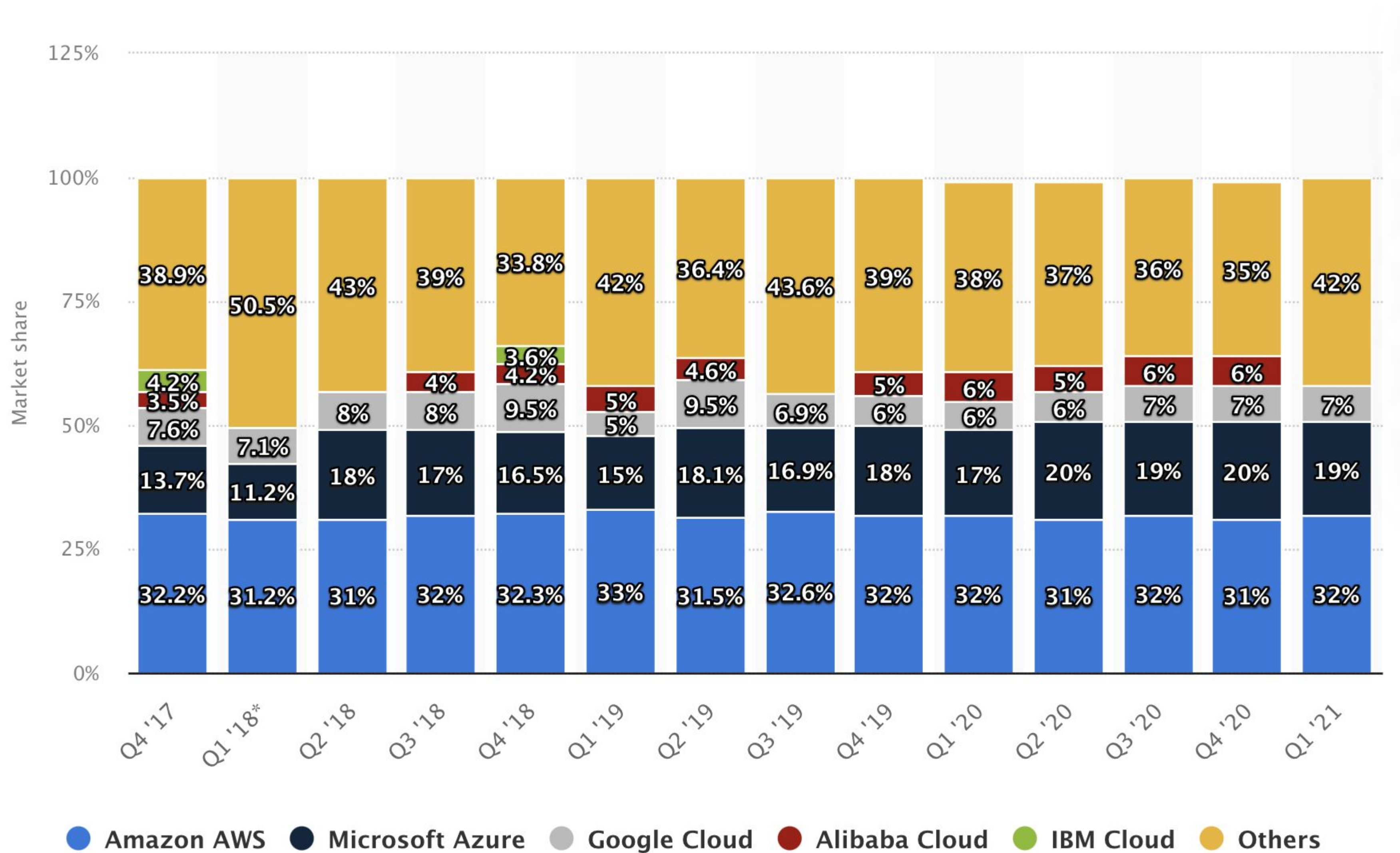
기타 중요 서비스 - AI & ML Services

- SageMaker
 - Deep Learning and Machine Learning end-to-end framework
- Lex
 - Conversational Interface (Chatbot service)
- Polly
 - Text to Speech Engine
- Rekognition
 - Image Recognition Service

기타 중요 서비스

- Amazon Alexa
 - Amazon's voice bot platform
- Amazon Connect
 - Amazon's Contact Center Solution
 - 콜센터 구현이 아주 쉬워짐
- Lambda
 - Event-driven, serverless computing engine
 - 서비스 구현을 위해서 EC2를 론치할 필요가 없음
 - Google Cloud에는 Cloud Function이란 이름으로 존재
 - Azure에는 Azure Function이란 이름으로 존재

글로벌 클라우드 시장 점유율



Redshift 소개

Redshift에 대해 더 자세히 알아보자

◆ Redshift: Scalable SQL 엔진 (1)

- ❖ 2 PB까지 지원

- ❖ Still OLAP

- 응답속도가 빠르지 않기 때문에 프로덕션 데이터베이스로 사용불가

- ❖ Columnar storage

- 컬럼별 압축이 가능
 - 컬럼을 추가하거나 삭제하는 것이 아주 빠름

◆ Redshift: Scalable SQL 엔진 (2)

❖ 벌크 업데이트 지원

- 레코드가 들어있는 파일을 S3로 복사 후 COPY 커맨드로 Redshift로 일괄 복사

❖ 고정 용량/비용 SQL 엔진

- vs. Snowflake vs. BigQuery

❖ 다른 데이터 웨어하우스처럼 primary key uniqueness를 보장하지 않음

- 프로덕션 데이터베이스들은 보장함

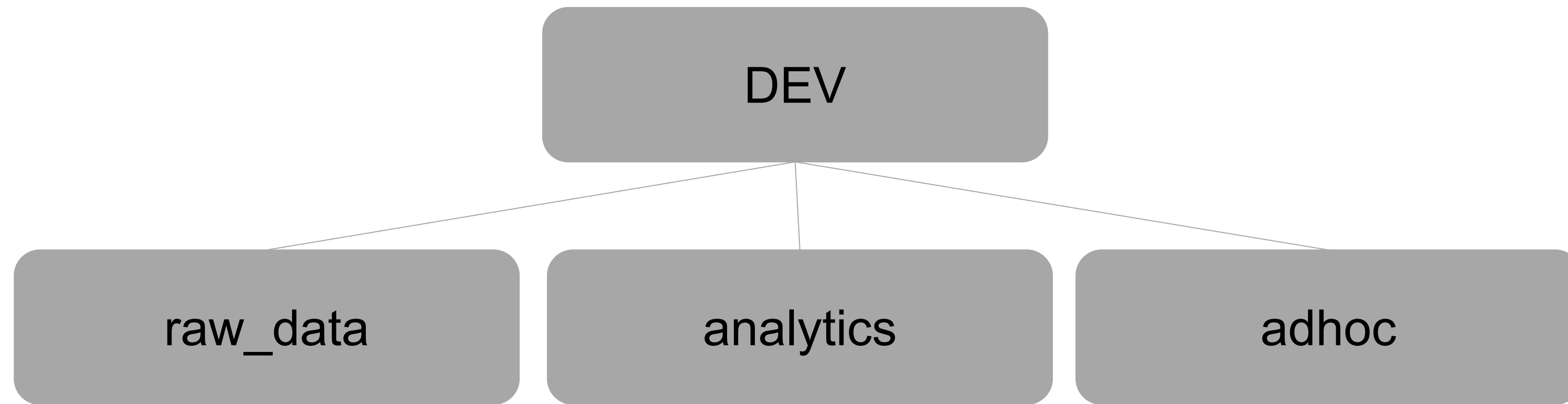
◆ Redshift는 Postgresql 8.x와 SQL이 호환됨

- ❖ 하지만 Postgresql 8.x의 모든 기능을 지원하지는 않음
 - 예를 들어 text 타입이 존재하지 않음
- ❖ Postgresql 8.x를 지원하는 툴이나 라이브러리로 액세스 가능
 - JDBC/ODBC
- ❖ 다시 한번 SQL이 메인 언어라는 점 명심
 - 그러기에 테이블 디자인이 아주 중요

Redshift Options and Pricing

	vCPU	Memory	Storage	Price	Monthly
Dense Storage					
ds2.xlarge	4	31	2TB HDD	\$0.85/hour	\$612
ds2.8xlarge	36	244	16TB HDD	\$6.80/hour	\$4,896
Dense Compute					
dc2.large	2	15	0.16TB SSD	\$0.25/hour	\$180
dc2.8xlarge	32	244	2.56TB SSD	\$4.80/hour	\$3,456
Managed Storage					
ra3.4xlarge	12	96	64TB SSD	\$3.26/hour	\$2,347
ra3.16xlarge	48	384	64TB SSD	\$13.04/hour	\$9,389

◆ Redshift Schema (폴더) 구성



```
CREATE SCHEMA raw_data;  
CREATE SCHEMA analytics;  
CREATE SCHEMA adhoc;
```



◆ Redshift 액세스 방법

- ❖ 이번 강좌에서는 Google Colab을 사용 예정
- ❖ Postgresql 8.x와 호환되는 모든 툴과 프로그래밍 언어를 통해 접근 가능
 - SQL Workbench (Mac과 윈도우), Postico (Mac)
 - Python이라면 psycopg2 모듈
 - 시각화/대시보드 툴이라면 Looker, Tableau, Power BI, Superset등에서 연결가능



퀴즈 풀어보기

<https://forms.gle/KR6SsNqZTpsr36Tc7>