

빠르고 손쉽게 클라우드 기반 데이터 분석 시작하기

정세웅

Data Analytics Solutions Architect, AWS

목차

- 데이터 분석 개요
- 데이터 분석 on AWS
- DEMO - 빠르고 쉽게 클라우드에서 분석 시작하기

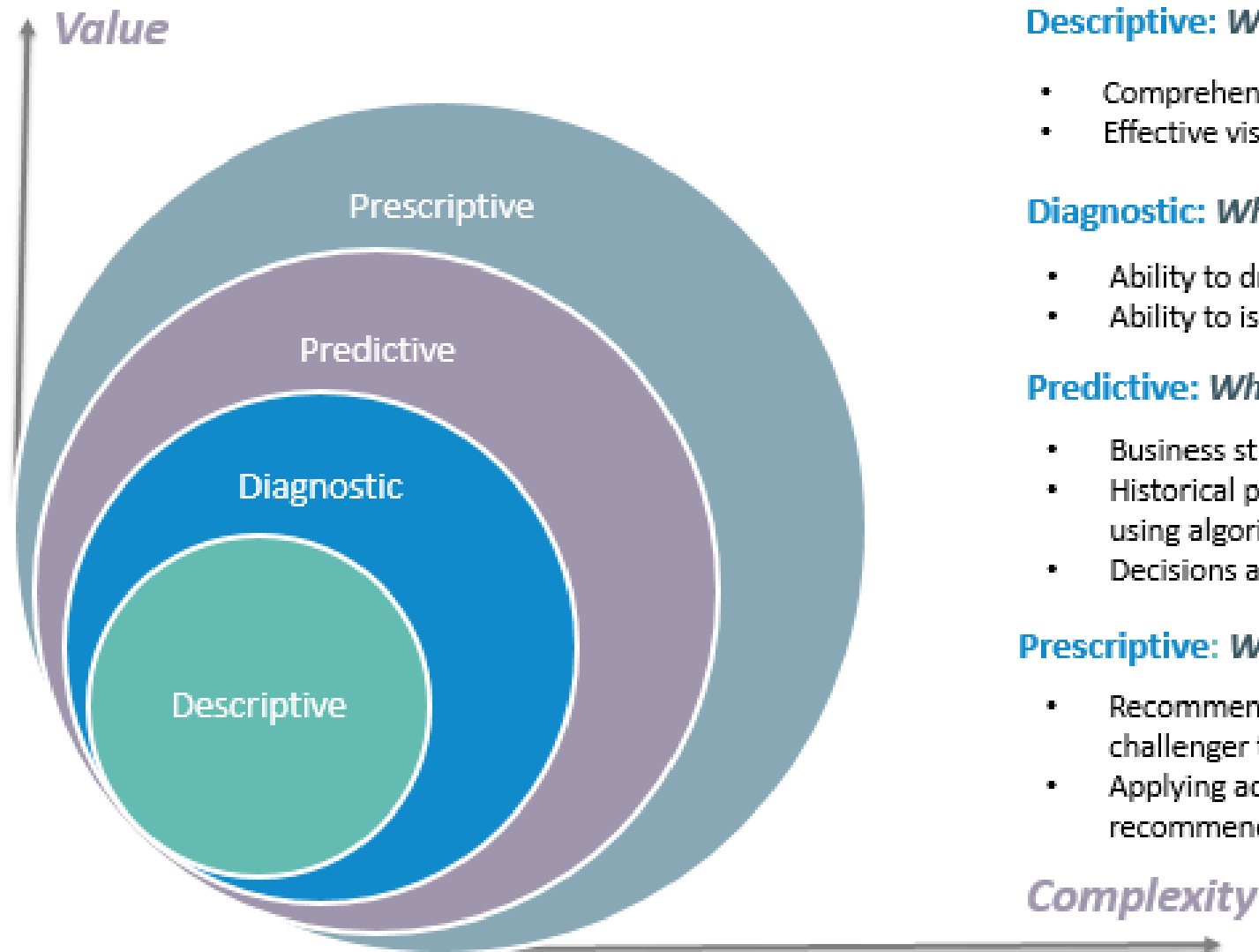
Data Analysis

Easiest way to analyze data

데이터 분석이란?

데이터 분석은 의사결정에 필요한 의미있는 정보와 근거들을 얻어내기 위해 데이터를 정리, 가공, 변환, 조사, 모델링 하는 과정을 뜻한다.(Wikipedia)

4 types of Data Analytics



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

데이터 분석의 목적

데이터 분석과 활용의 가장 기본적인 목표는 비즈니스의 성공

고객 만족을 우선으로

- 고객 Behavior 분석
- 고객 분류
- 추천 / 개인화

매출을 증가시키고

- 매출 추이 / 퍼널 분석
- 전환률, 리텐션 향상

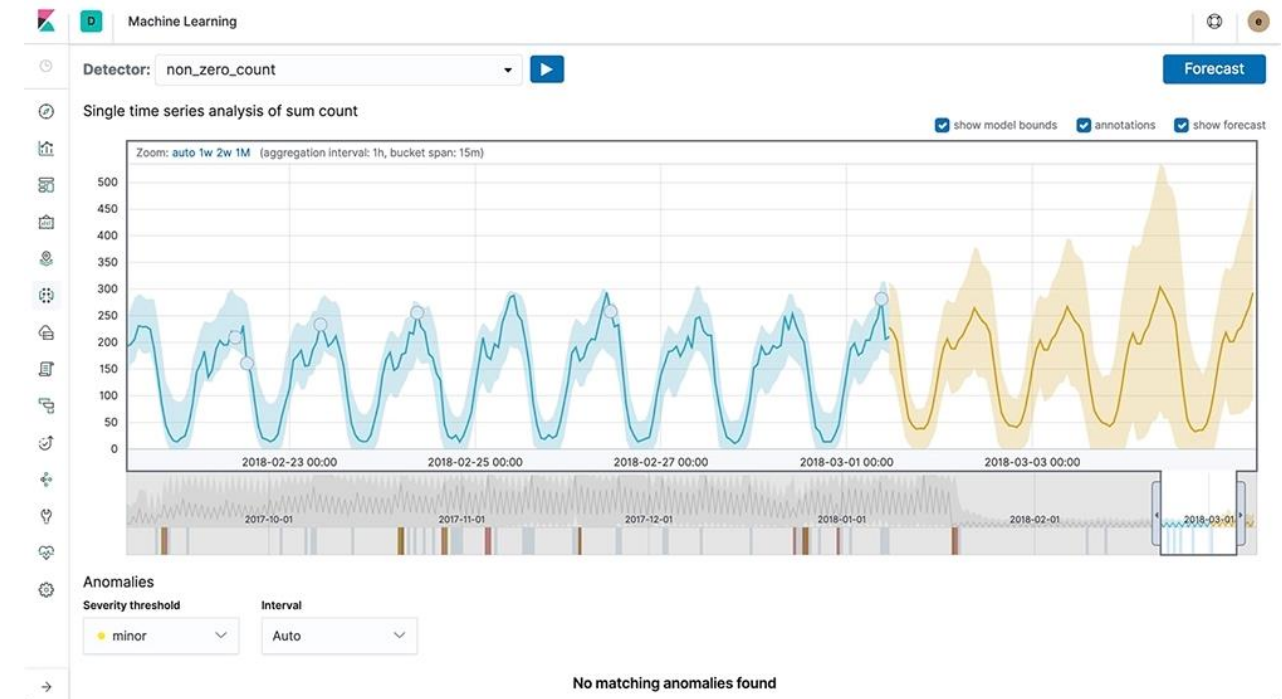
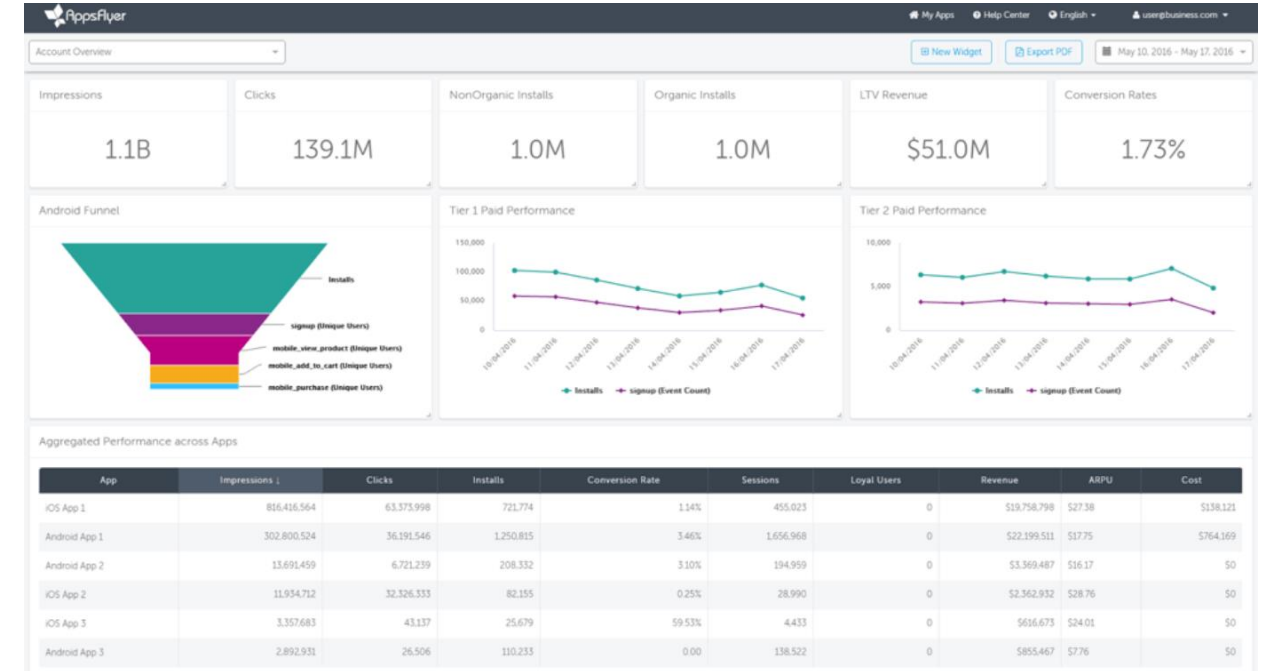
비용을 감소 시킨다.

- 마케팅 성과 측정
- Fraud detection, Risk analysis

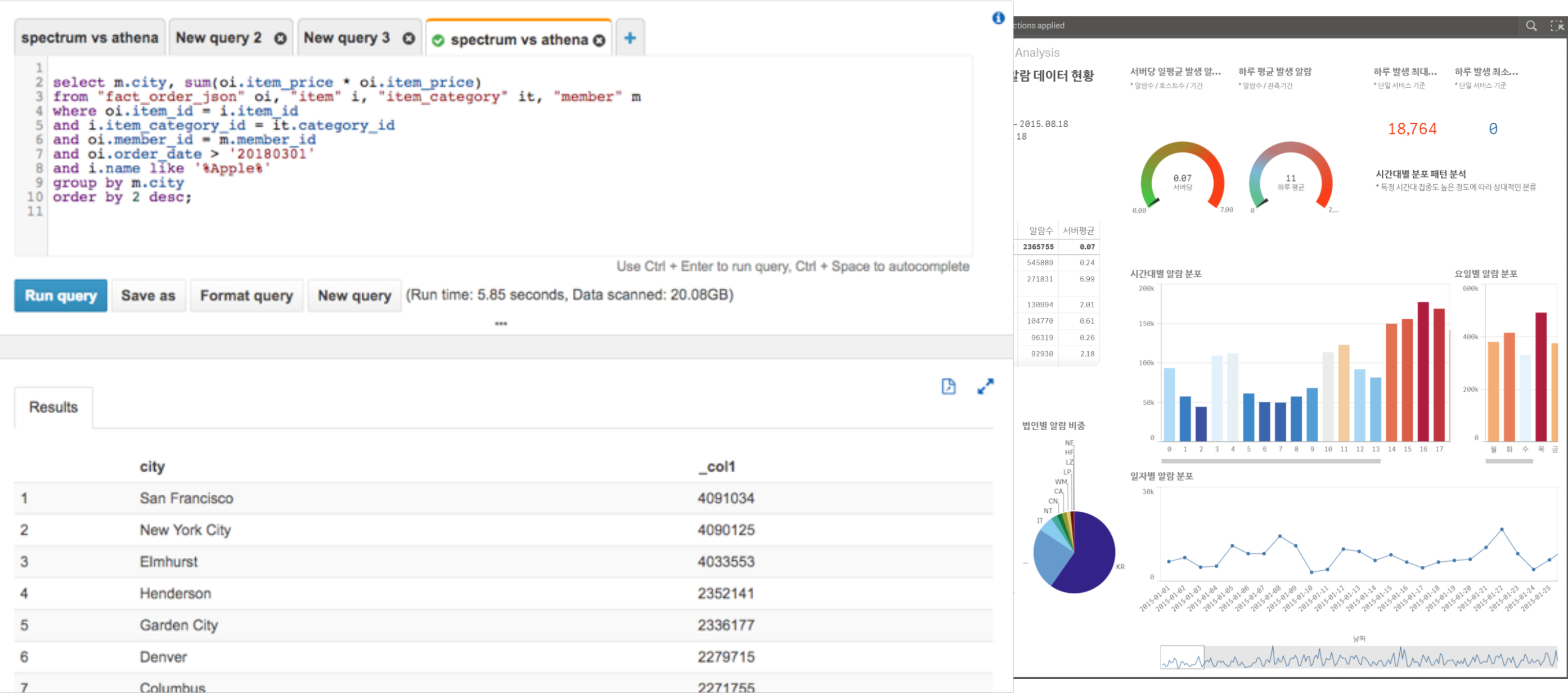


데이터 분석의 주요 결과물

- 매출 지표, 예측 수치
- 고객 유형, Behavior 분석 결과
- 퍼널 분석 결과 (이탈/잔존/전환율)
- 마케팅 성과 지표
- 추천 / 개인화 모델
- 사기, 리스크 탐지 모델



데이터 분석의 결과물 공유



데이터 분석의 결과물 공유

File Edit View Run Kernel Git Tabs Settings Help

1. Data Transform with Glue X 2-1. R - install_r_kernel.ipyr X

Code

이후 데이터 변환 작업의 편의성을 위해서 Glue의 DynamicFrame

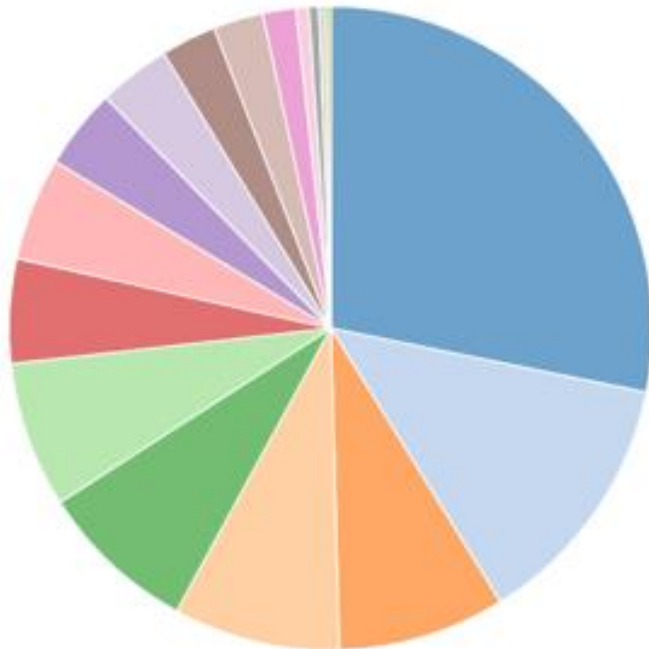
```
[2]: order = glueContext.create_dynamic_frame.from_
      print "Count: ", order.count()
      order.printSchema()
      order.show(5)
```

Count: 11283758
root
|-- member_id: string (nullable = true)
|-- order_date: long (nullable = true)
|-- order_status: string (nullable = true)
|-- country: string (nullable = true)
|-- shipping_date: string (nullable = true)
|-- total_price: long (nullable = true)
|-- city: string (nullable = true)
|-- order_time: long (nullable = true)
|-- state: string (nullable = true)
|-- postal_code: long (nullable = true)
|-- region: string (nullable = true)
|-- order_id: string (nullable = true)

member_id	order_date	order_status	count_id
ND-18370	20161119	shipped	United Stat 778
KH-16330	20170619	shipped	United Stat 944
SP-20860	20150816	shipped	United Stat 753

어떤 상품 카테고리의 매출이 높은지 확인 - Machir
Phones, Copiers 순서로 매출이 나타남

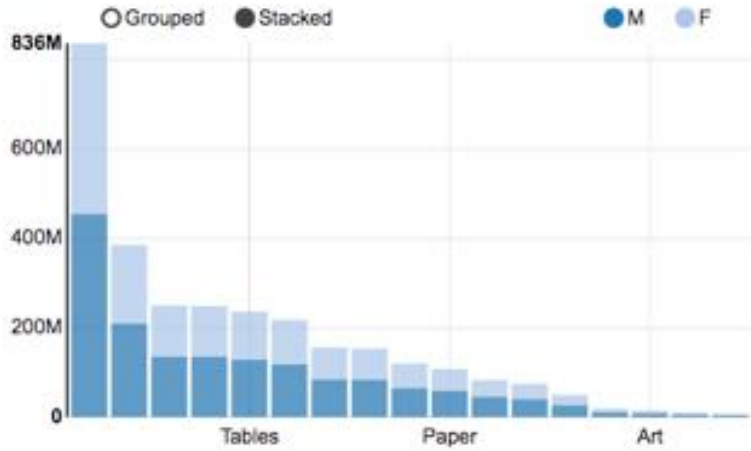
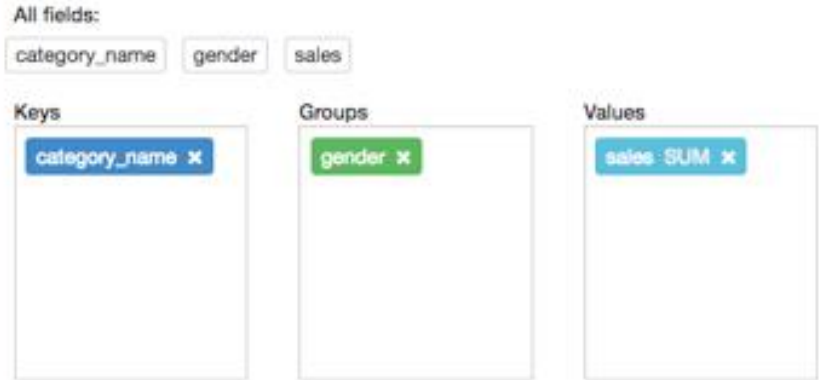
```
%sql
select item_category.category_name, sum(order_item.item_count
*order_item.item_price) sales
from order_item, item, item_category
where order_item.item_id = item.item_id
and item.item_category_id = item_category.category_id
group by item_category.category_name
order by sales desc
```



Took 22 sec. Last updated by anonymous at May 03 2018, 1:49:32 PM. (outdated)

카테고리 매출 비중을 성별에 따라 분류해 보자

```
%sql
select item_category.category_name, member.gender, sum(order_item
.item_count*order_item.item_price) sales
from order, order_item, item, item_category, member
where order_item.item_id = item.item_id
and item.item_category_id = item_category.category_id
and order.order_id = order_item.order_id
and order.member_id = member.member_id
group by item_category.category_name, member.gender
order by sales desc
```



Took 1 min 31 sec. Last updated by anonymous at May 03 2018, 3:09:42 PM. (outdated)

데이터 분석의 결과물 공유

Amazon SageMaker

Overview Hide

Notebook instance

Explore AWS data in your notebooks, and use algorithms to create models via training jobs.

Create notebook instance

Amazon SageMaker Studio

File Edit View Run Kernel Git Tabs Settings Help

xgboost_customer_churn.ipynr

conda_amazonei_mxnet_p27

- Have the predictor variable in the first column
- Not have a header row

But first, let's convert our categorical features into numeric features.

```
[ ]: model_data = pd.get_dummies(churn)
      model_data = pd.concat([model_data['Churn?_True'], model_data.drop(['Churn?_True'], axis=1)], axis=1)
```

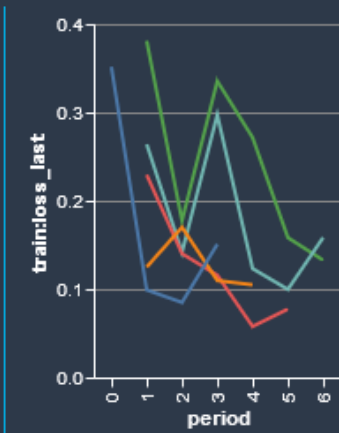
And now let's split the data into training, validation, and test sets. This will help prevent us from overfitting the model, and allow us to test the models accuracy on data it hasn't already seen.

```
[ ]: train_data, validation_data, test_data = np.split(model_data.sample(frac=1, random_state=123), [int(len(model_data) * 0.7), int(len(model_data) * 0.8)])
      train_data.to_csv('train.csv', header=False, index=False)
      validation_data.to_csv('validation.csv', header=False, index=False)
```

Now we'll upload these files to S3.

```
[ ]: boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'train.csv')).upload_file(train_data.to_csv(index=False).get_value())
      boto3.Session().resource('s3').Bucket(bucket).Object(os.path.join(prefix, 'validation.csv')).upload_file(validation_data.to_csv(index=False).get_value())
```

Trial Component Chart



Trial Component List

10 rows selected

Add chart

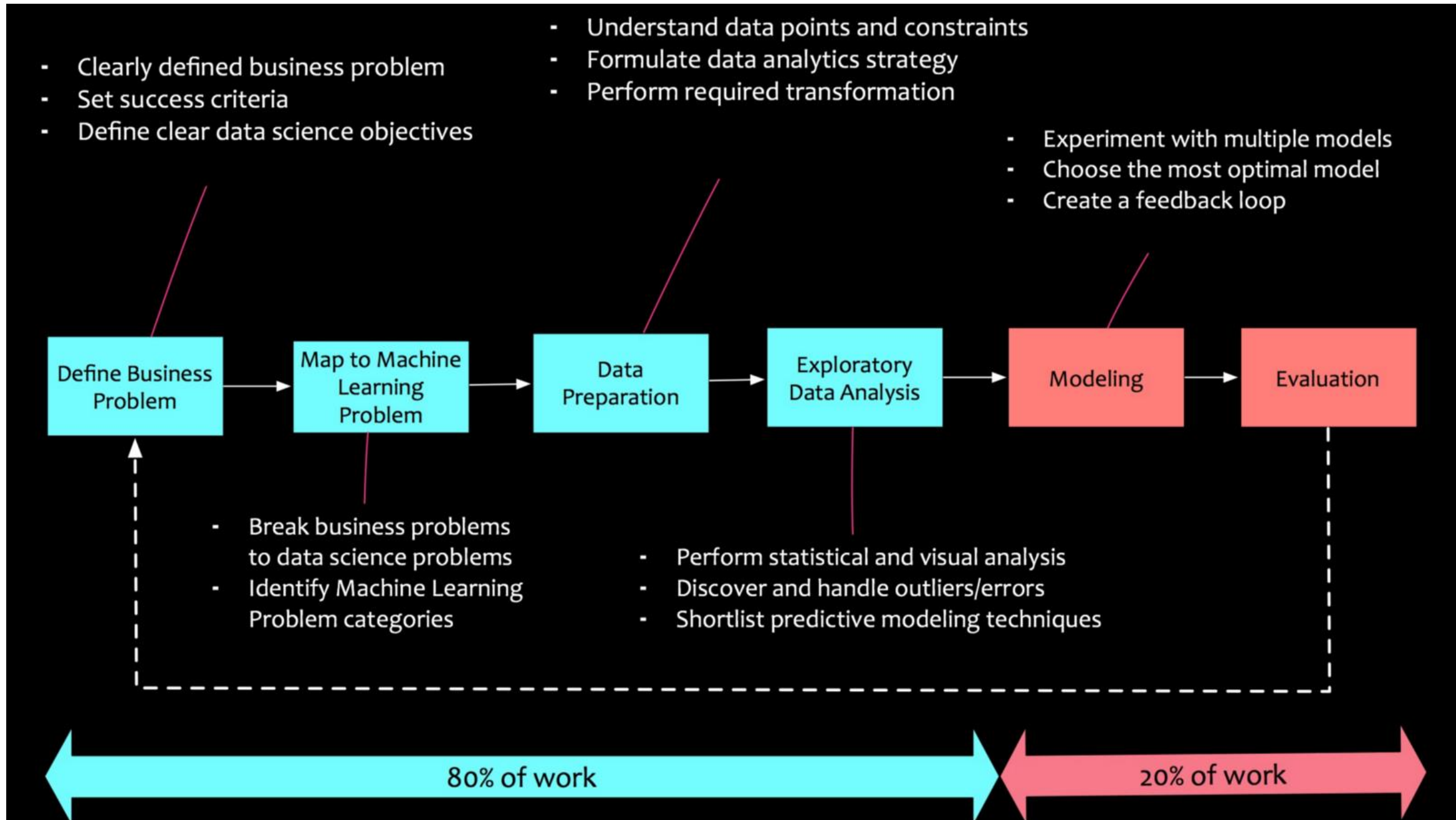
Deploy model

Status	Experiment	Type	Trial	Trial c
✓ Completed	customer-churn-predi...	Training job	Trial-3	Tra
✓ Completed	customer-churn-predi...	Training job	Trial-2	Tra
✓ Completed	customer-churn-predi...	Training job	Trial-1	Tra
✓ Completed	customer-churn-predi...	Training job	Trial-0	Tra

0 2 conda_amazonei_mxnet_p27 | Idle

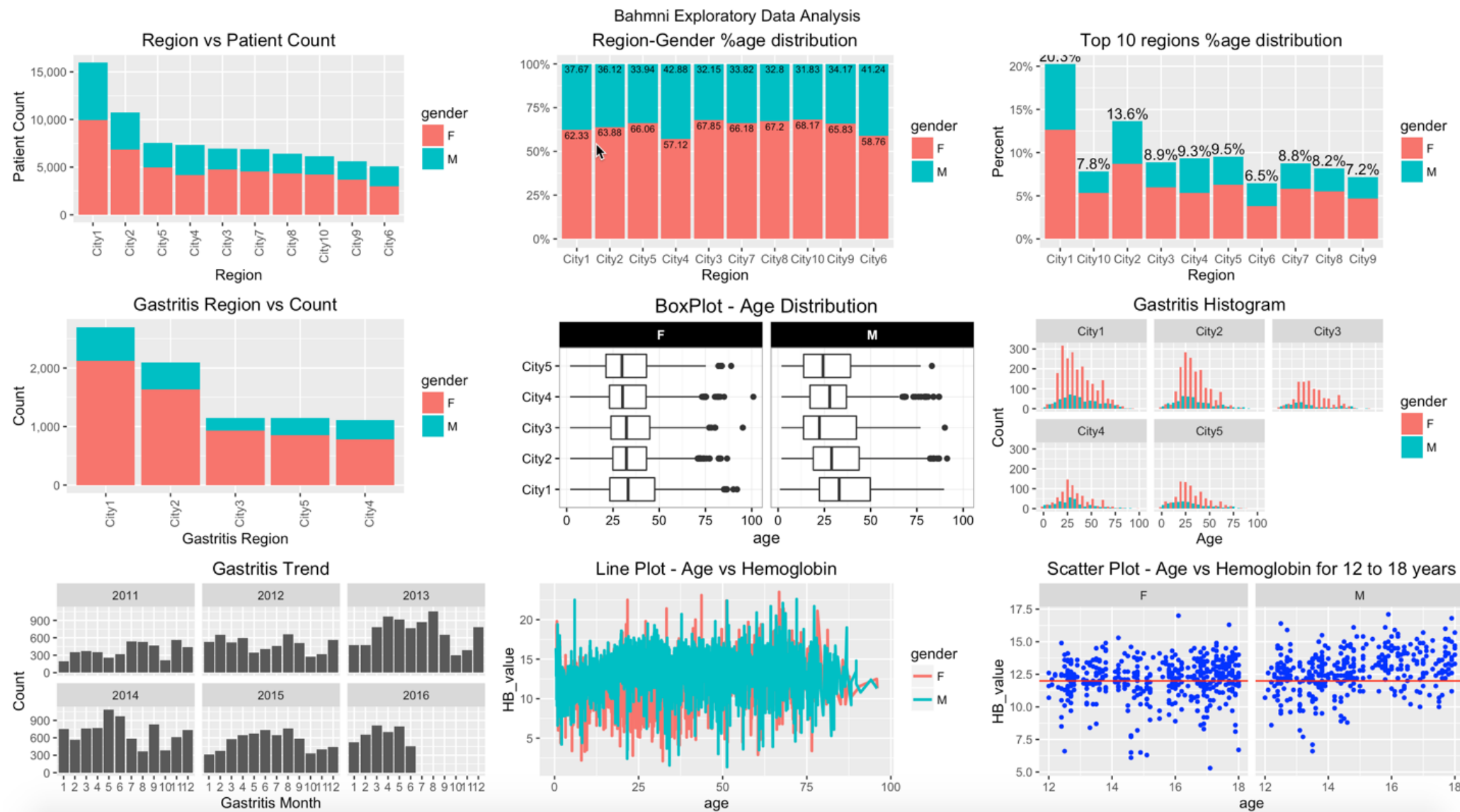
Mode: Command Ln 1, Col 1 xgboost_customer_churn.ipynb

데이터 분석의 주요 과정



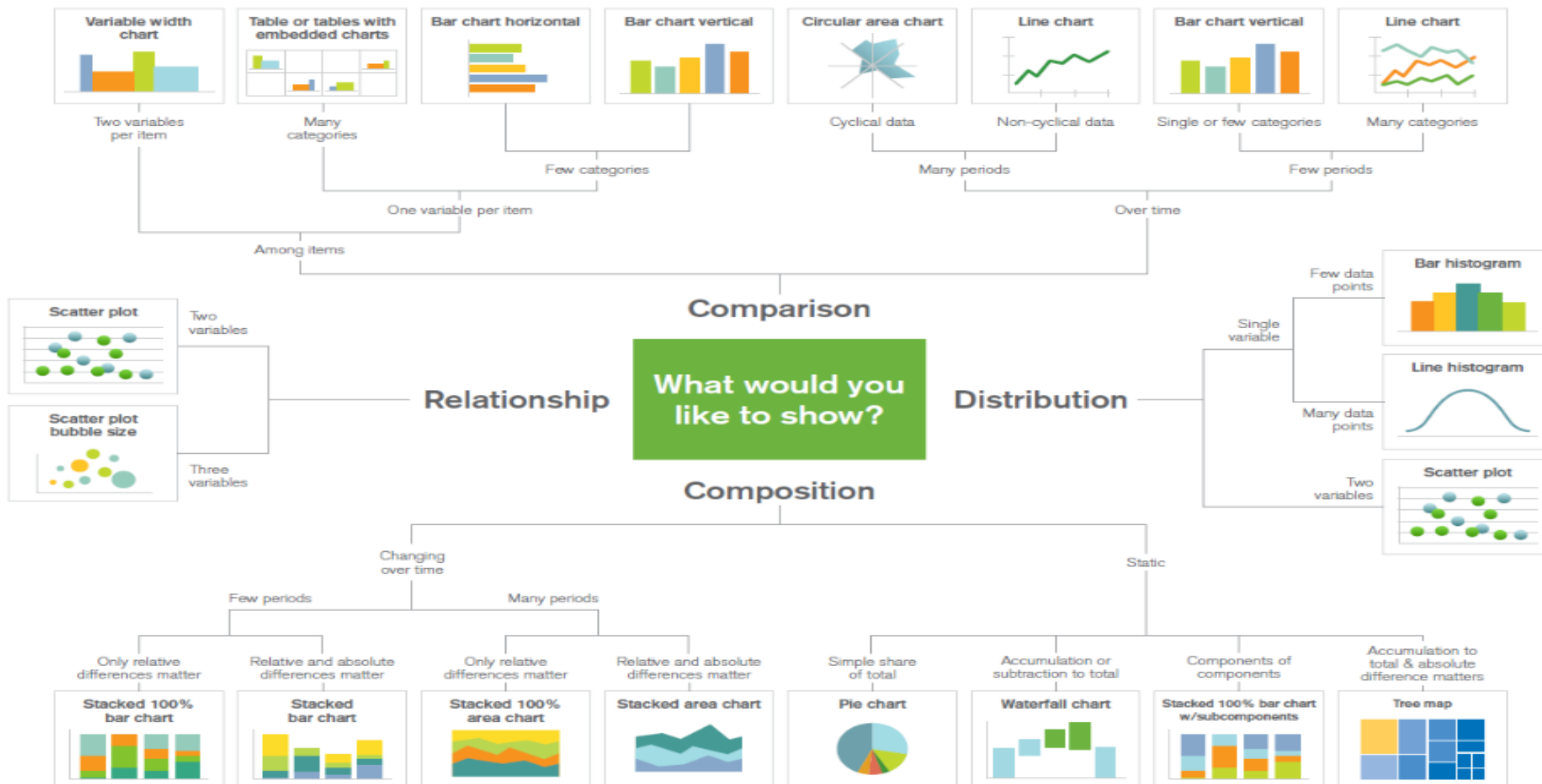
EDA(Exploratory data analysis)

EDA(탐험적 데이터 분석)은 데이터 분석의 초기 단계에 이해하기 위해 다양한 차트를 이용하여 데이터를 표현하고 이해하는 과정이다. 이를 통해 문제 해결에 필요한 데이터를 식별하고, 가설 수립, 알고리즘 선정 등을 할 수 있다.



데이터 시각화

분포, 분류, 관계, 비교, 시간, 공간



데이터 분석과 엔지니어링

Prepare data pipeline, data catalog, and ETL for data analysis

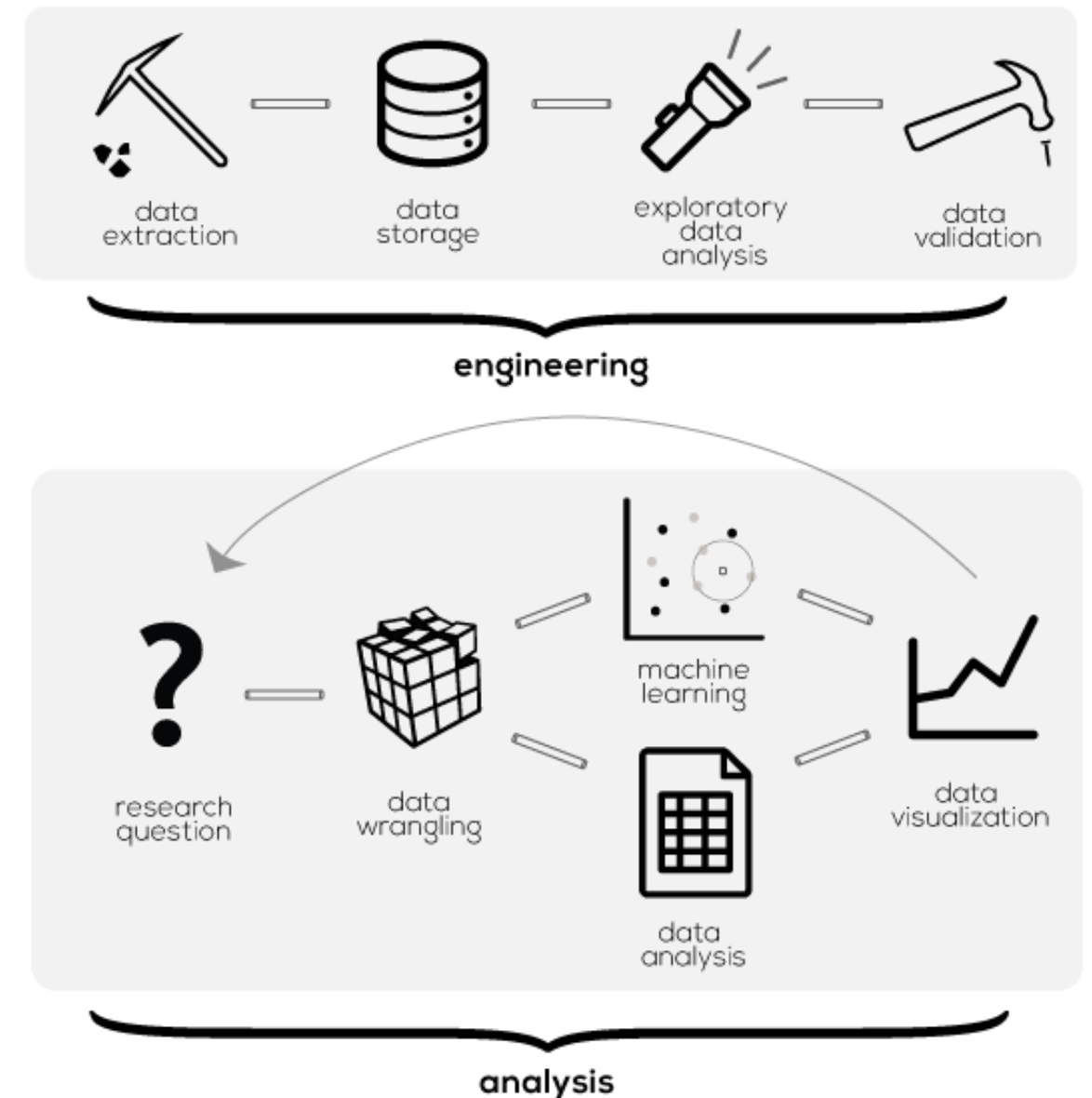
데이터 분석을 위한 파이프라인

- 데이터 식별 및 추출
- 데이터의 이동 / 저장
- 데이터의 탐색 / 분석
- 수집 / 활용의 자동화

데이터 분석의 활용

- 가설을 수립 / 검증
- 데이터의 변환 / 처리
- 통계 / ML 기반 모델링
- 시각화 / 결과 데이터 제공

the data pipeline



Data Analytic on AWS

Easiest way to analyze data

분석 플랫폼의 데이터 파이프라인



수집



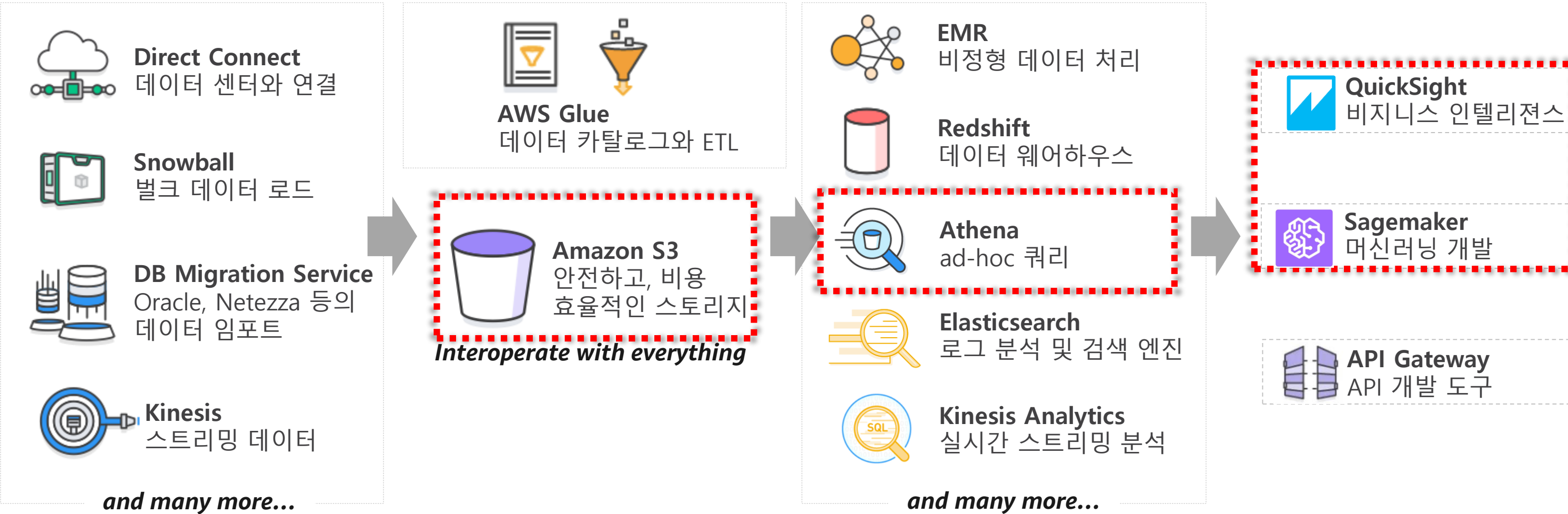
저장



처리 및 분석



소비

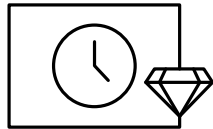


높은 확장성과 내구성을 가진 데이터레이크 - Amazon S3

밀리세컨드 내의 지연시간을 가지는 보안성과 높은 확장성, 내구성을 가지는 오브젝트 스토리지

모든 유형의 데이터 저장소 - 웹사이트, 모바일앱, 기업용 어플리케이션, IoT센서

내구성, 가용성과 확장성



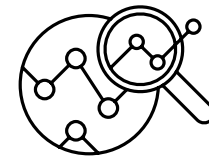
99.999999999%의 내구성 설계; 데이터는 **AWS Region** 내부에 3곳의 물리적 공간에 분산 저장됨; 자동으로 다른 **AWS Region** 에 복제 구성 가능

보안과 컴플라이언스



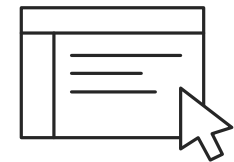
3가지의 다른 형태의 암호화기능을 제공; 리전 간 전송시에도 데이터 암호화 전송; **CloudTrail** 로 로그 및 모니터링하며, **ML** 기반 **Macie** 로 민감 데이터를 찾아내고 보호할 수 있음

즉각적인 쿼리수행



데이터 이전 없이 **DataLake** 에서 분석 및 **ML** 을 실행할 수 있음; **S3 Select** 를 사용해서 데이터의 하위 집합을 검색하고 분석 퍼포먼스를 400% 증가할 수 있음

유연한 관리



데이터 사용 트렌드를 분류,보고 및 시각화; 오브젝트에 태그를 붙여 스토리지 사용과 비용 및 보안을 확인가능; 보관기간 및 **Tiering** 을 자동화하는 수명주기관리 정책 작성

서버리스 Ad-hoc 쿼리 엔진 - Amazon Athena

표준SQL을 사용해서 Amazon S3의 데이터를 분석하는 대화식 쿼리 서비스
설정 및 관리해야 할 인프라도 없으며, 로드 해야할 데이터도 없음

Amazon Glacier 에 보관된 데이터에 대해 SQL 쿼리를 실행할 수 있음

즉각적인 Query



셋업 비용이 들지 않음;
S3를 바로 지정하고
쿼리를 수행하면 됨

Query 당 비용



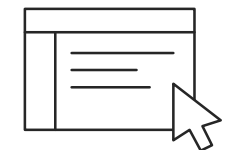
쿼리 실행에 대해서만
지불; 압축을 통해서
쿼리당 **30-90%** 비용
절감 가능

개방



ANSI SQL 인터페이스,
JDBC/ODBC 드라이버,
다양한 포맷, 압축
유형, 복잡한 조인 및
데이터 타입

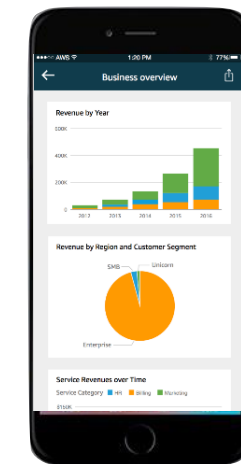
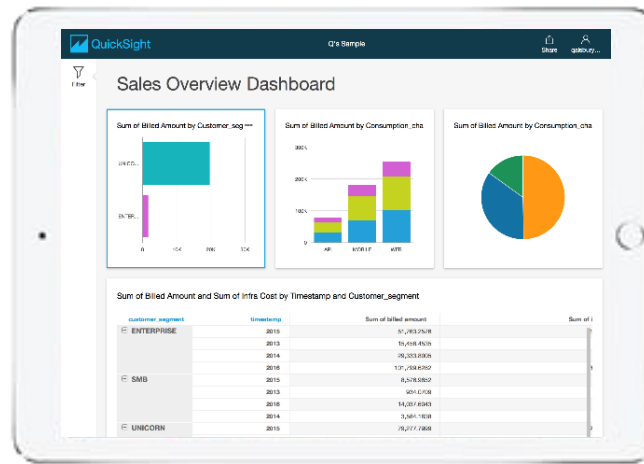
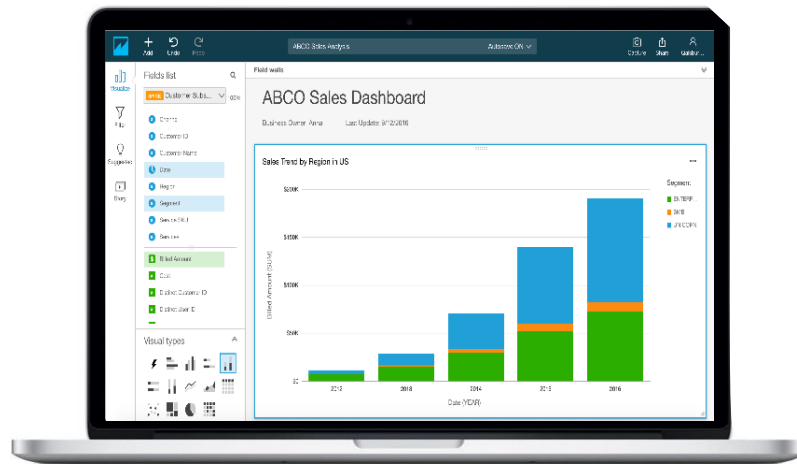
손쉬운 사용



서버리스: 인프라 없음,
관리 불필요
QuickSight 와 통합

분석, 협업, 시각화 대시보드 - Amazon QuickSight

QuickSight를 통해 사용자는 대시보드를 통해 쉽게 데이터와 분석 결과를 공유할 수 있으며, 다양한 디바이스에서 스토리 보드에 접근 가능



분석

분석에서 데이터를 시각적으로 탐색하는 것은 매우 중요합니다. 사용자가 쉽게 다양한 방식으로 시각화 할 수 있고, 협업 할 수 있도록 도와줍니다.

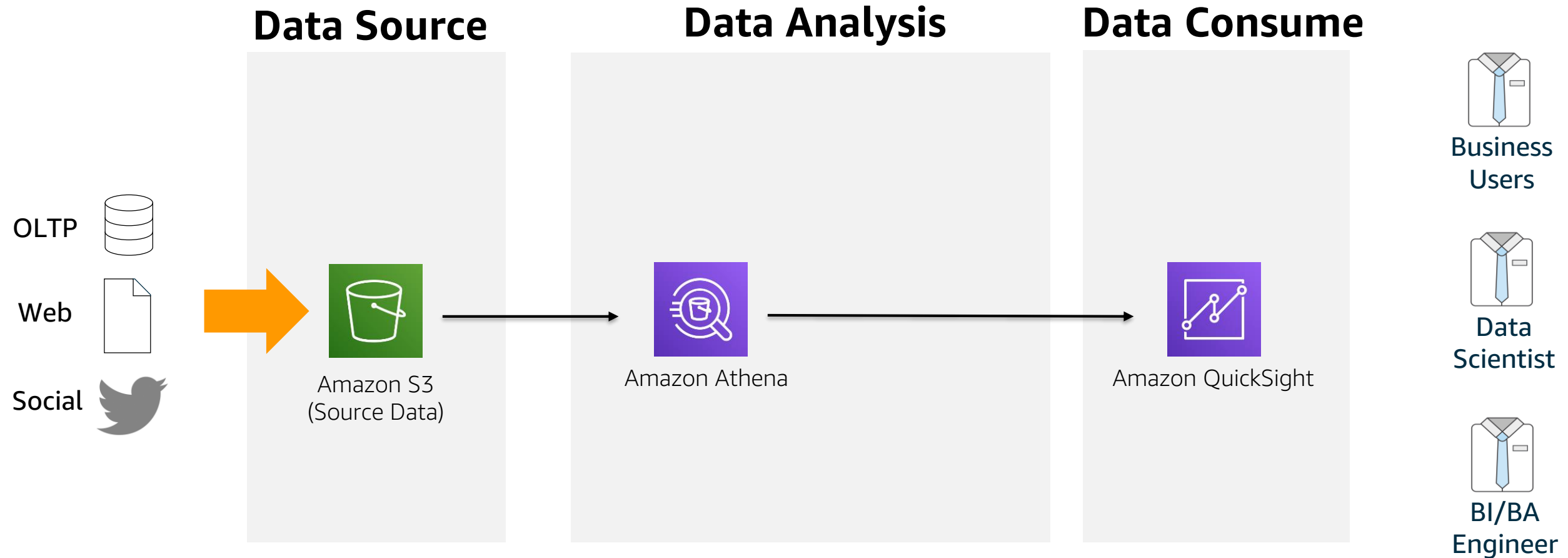
대시보드

여러분의 분석 결과물을 대시보드 형태로 공유 할 수 있습니다. 뷰어 모드로 사용하는 고객에게는 세션별 과금 정책으로 매우 저렴하게 사용 가능합니다.

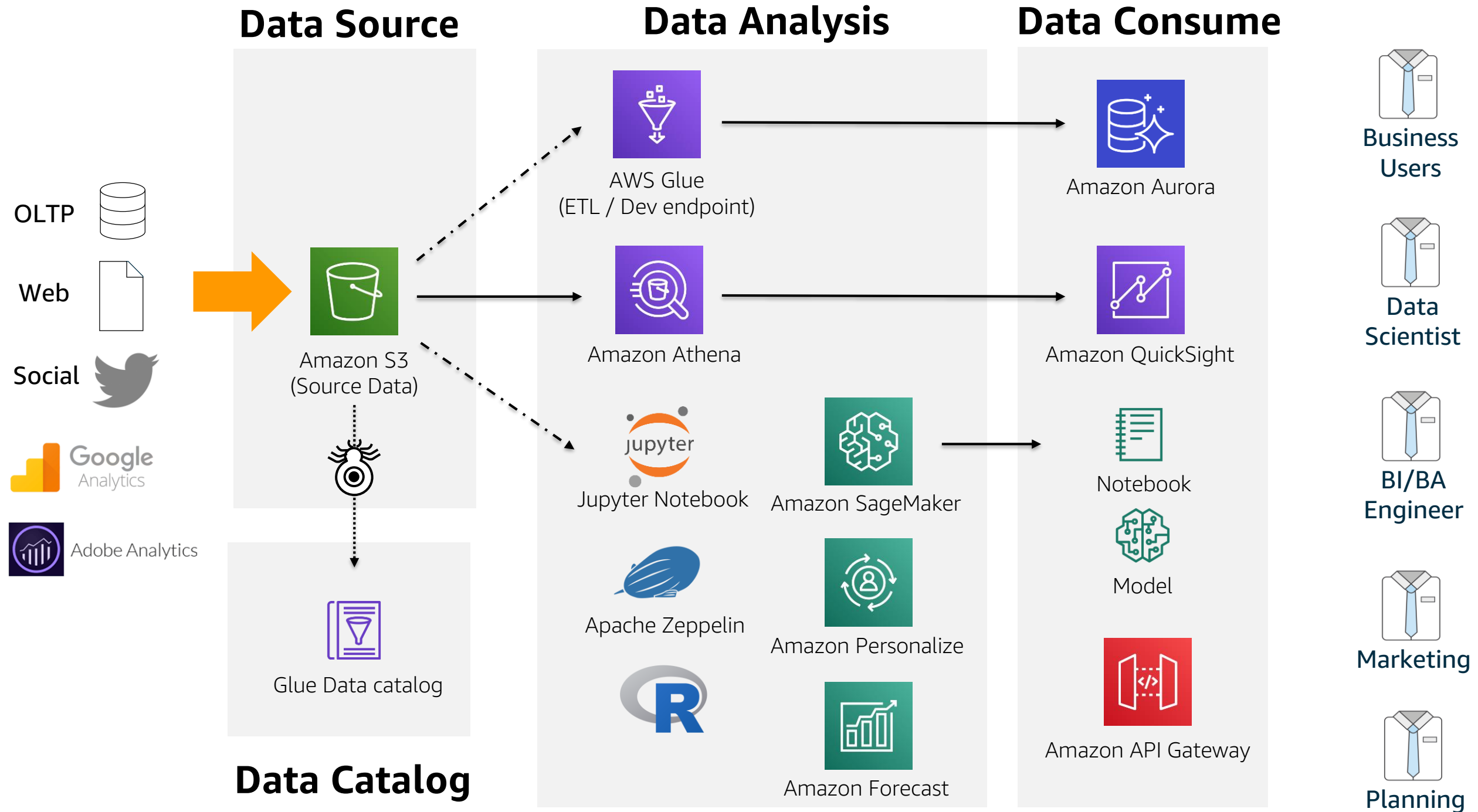
스토리보드

디바이스에 상관없이 스토리 보드를 통해 분석 결과를 공유하세요

Easy Data Analysis 파이프라인

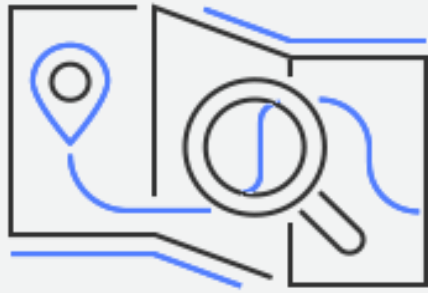


Data Analysis 파이프라인의 확장



DEMO - 빠르고 쉽게 클라우드에서 데이터 분석 시작하기

AWS 교육 및 자격증



조직을 위한
맞춤 교육

고객 및 파트너를 위해
준비된 데이터 및
데이터베이스 관련
맞춤형 교육 여정을
확인해 보세요.



원하는 방법으로
- 유연한 학습 형태

“The elements of Data
Science” 과정을 포함한
무료 디지털 교육 또는
강의실 교육을 통해
클라우드 역량을
향상시키세요.



AWS 자격증을 통한
기술 역량 입증

업계에서 인정받는
데이터 분석 또는
데이터베이스 - 전문분야
AWS 자격증을 통해
전문성을 입증할 수 있습니다.

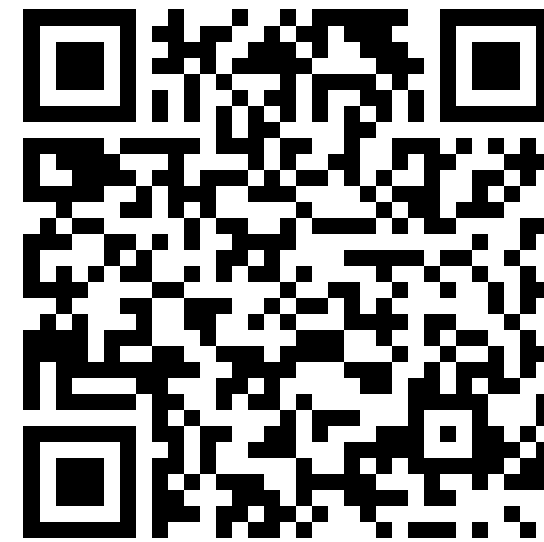
aws.amazon.com/training

AWS 데이터분석 관련 자료를 원하시면 ...

데이터 분석 관련 기술 백서 및 전자책을 자세히 살펴보면
데이터에서 새로운 통찰력과 가치를 발견 할 수 있습니다!

- 클라우드 기반 데이터 분석 서비스
- 데이터의 분석 활용 사례
- 최신 분석 아키텍처 생성 방법
- 데이터 중심 기업 전환
- 동영상, 기술 백서 등

지금 방문하세요! »



<https://tinyurl.com/data-databases-analytics-kr>

AWS 데이터 분석 특집 웨비나에 참석해주셔서 대단히 감사합니다.

저희가 준비한 내용, 어떻게 보셨나요?
더 나은 세미나를 위하여 **설문을 꼭 작성해 주시기 바랍니다.**



aws-korea-marketing@amazon.com



twitter.com/AWSKorea



facebook.com/amazonwebservices.ko



youtube.com/user/AWSKorea



slideshare.net/awskorea



twitch.tv/aws

Thank you!