AWS re:Invent

FSI305

Barclays: Accelerating research data science using AWS

Vojtech Kral
Data Science Platform Engineer
Barclays

William McWeeny
Lead Architect
Barclays

Rob Charlton
Principal Solutions Architect, Financial Services
Amazon Web Services

Introduction: Session

Why data science?

Creating a research data science platform using AWS services

Benefits of this solution include the ability to

- Securely import large amounts of data
- Provision resources to quickly execute models
- Extract insights at a greater velocity

However, to realize these benefits, a number of challenges needed to be overcome

In this session, we will share how we overcame these challenges

Introduction: Tensions and challenges

Tension

- Agility and control & security

Challenge

- Increasing agility while maintaining a controlled environment

Exacerbated

- By working in a regulated environment

What follows is

- A description of four challenges familiar to those working with this tension
- Solutions we found to overcome these challenges
- Our insights, learning, and “tips for the top”

Introduction: Context

Research business

Importance of data

Research data platform

- Datasets
  - Traditional financial data sources
  - New/alternative data sources
- Analysis & experimentation
- Batch job capabilities

Challenge 1: Downloading large external datasets

Challenge 1: Downloading large external datasets

New alternative datasets

Typically semi-structured data, 10 - 100s TiB

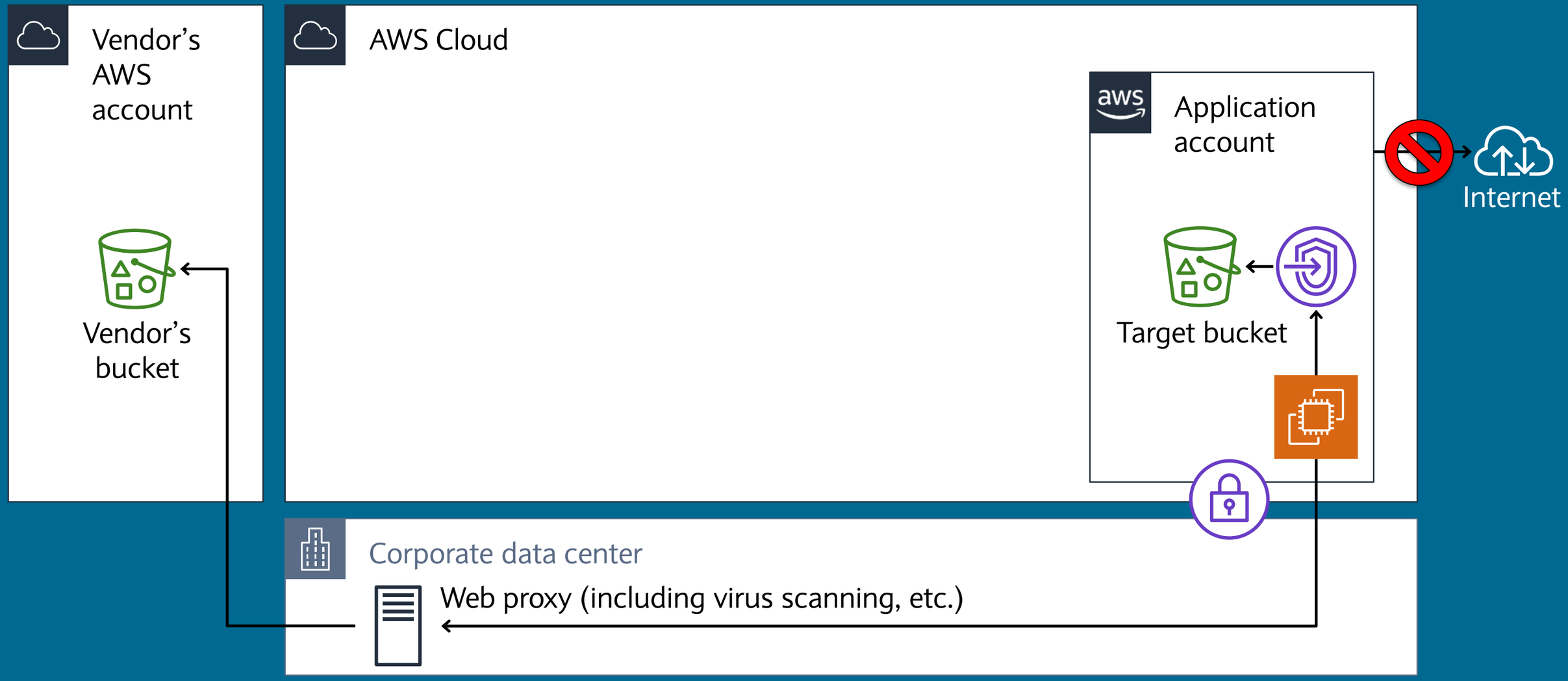
On-premises challenges

- Storage
- Bandwidth (for example, 100 TiB @ 1 Gbit/s = ~10 days)

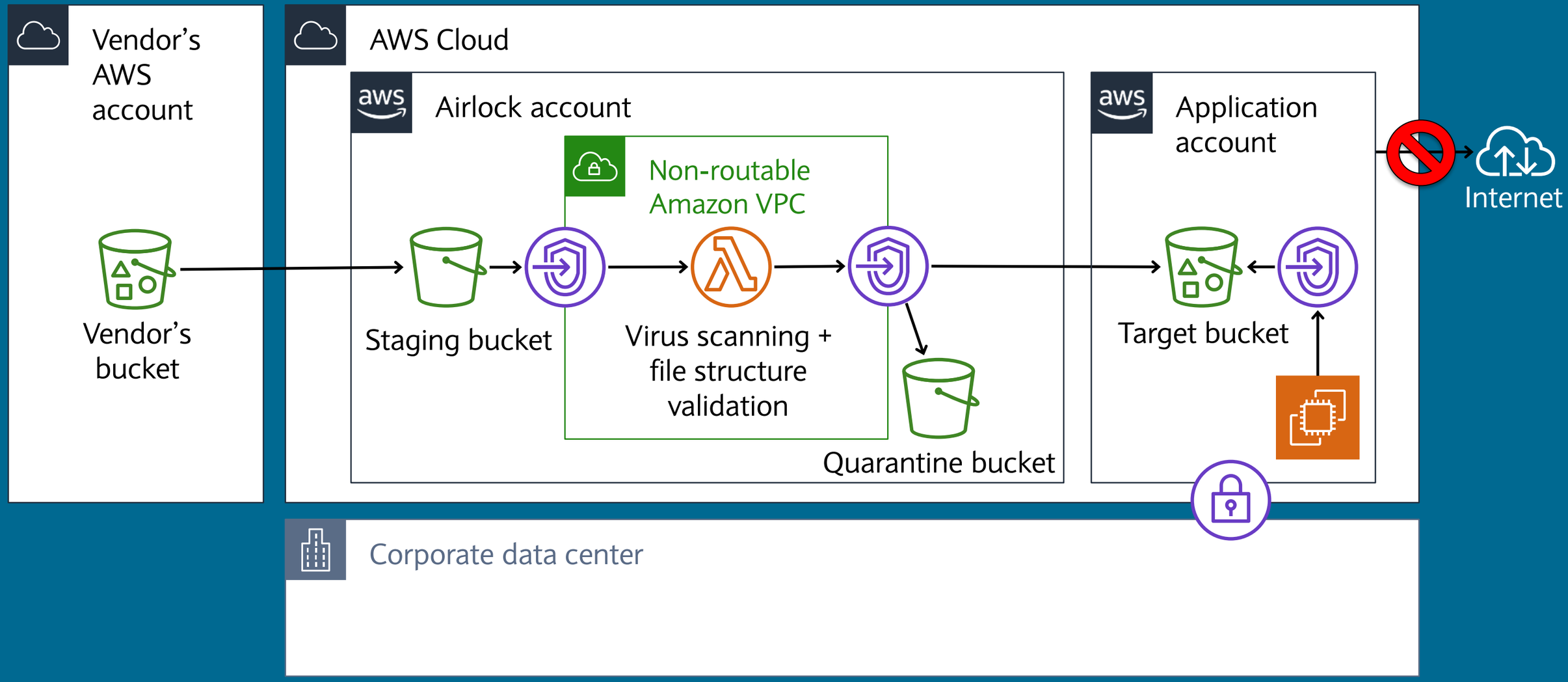
Delivery requirements

- Control
- Speed
- Minimal change

Challenge 1: Original architecture: Web proxy



Challenge 1: Architecture (Solution): Airlock account



BARCLAYS

Challenge 1: Permissions (Solution): Invest to save

Policies that are easier to read

- Are quicker to review
- Lead to fewer mistakes
- Are easier to maintain

Implicit deny approach = harder to read

- Requires evaluation of the whole resource policy and all users & roles

Explicit deny approach = easier to read

- Define maximum scope in a single place
- Safe addition of new users & roles

Challenge 1: Permissions (Solution): Deny wildcards

Challenge: Using partial wildcards with deny statements

```
{ "Effect": "Deny",
  "NotPrincipal": {
    "AWS": ["arn:aws:sts::444455556666:assumed-role/some-role/*" ]
} } // This is invalid - NotPrincipal does not support wildcards
```

Solution: “Deny all” with a condition to reduce scope

```
{ "Effect": "Deny",
  "Principal": "*",
  "Condition": {
    "StringNotLike": {
      "aws:userid": ["SOMEROLEID:*", "444455556666"] }
} }
```

Ref: https://aws.amazon.com/premiumsupport/knowledge-center/explicit-deny-principal-elements-s3/

Challenge 1: Requirements delivered

Control

- Approved reusable pattern providing full scanning requirements
- Separate account used to enforce segregation of duties

Speed

- Full end-to-end transfer rate peaking at ~3 GiB/s (~25 Gbit/s)

Minimal change

- Addition of new account resulting in few changes to controls in existing account

Challenge 1: Lessons learned

Complex human systems versus complicated problems

- Technology
- Processes/standards
- People


Security


Infrastructure


Application

For critical controls, consider using explicit deny within resource policies

- Improves security (and robustness to change), but also
- Saves time in security reviews

Challenge 2: Lead time for on-premises infrastructure

Challenge 2: Lead time for on-premises infrastructure

Primary challenge

- Business requirement to expand dataset would have exhausted storage capacity
- Lead time for additional storage did not meet project timelines
- Also, storage increase dependant on server upgrade (additional delay)

Solution

- Migrate on-premises relational database to Amazon Relational Database Service (Amazon RDS)

Outcome

- Project delivered quicker than would have been possible on-premises
- Amazon RDS decouples compute, storage, and database—almost eliminates upgrade effort

Challenge 2: Lessons learned

Lift and shift is a valuable migration pattern

- Faster time to market

However, please note, some refactoring is still required

- For example, not all database operations are supported by Amazon RDS

Hybrid hosting introduces greater complexity and possible additional work

- For example, when working with secrets management and/or database account maintenance

Consider migrating all tiers at once

Challenge 3: Need for a flexible compute environment

Challenge 3: Need for a flexible compute environment

Tooling for big data analysis

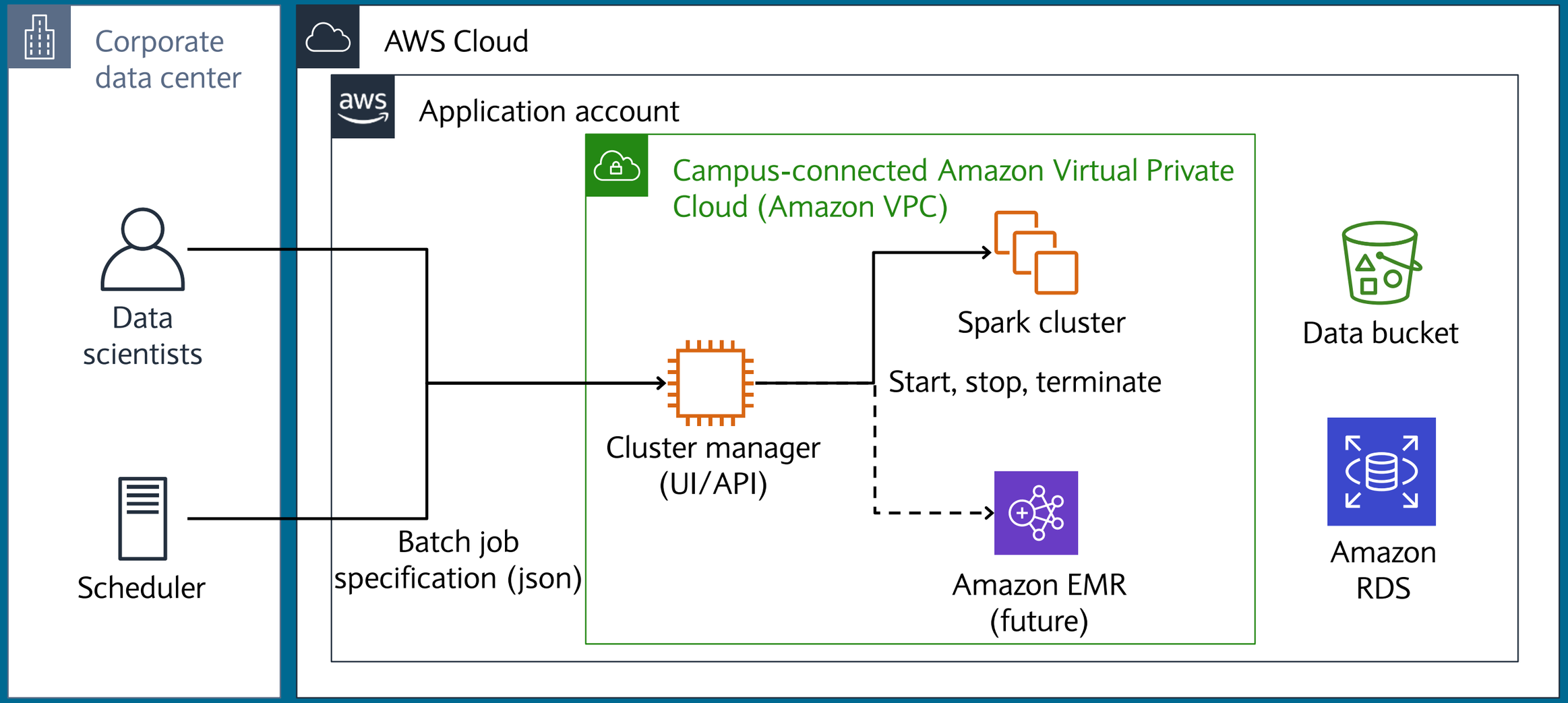
Platform selection

- Target: Amazon EMR
- Immediate: Custom Spark cluster with Jupyter Notebook

Single tenant clusters

- Flexibility
- No contention

Challenge 3: Solution architecture



Challenge 3: Outcome

Timely delivery

Elastic scale

- 5,000 vCPU, 20 TB RAM for a few hours

Single tenant clusters

- Flexibility, no contention

Challenge 3: Lessons learned

Deliver what is achievable but know your target state

Abstraction layer

- Easy iteration
- Custom features
- Greater control over permissions

Single-tenant cluster can be cheaper

- You know when to stop it
- Right sizing

Challenge 4: Better infrastructure utilization

Challenge 4: Better infrastructure utilization

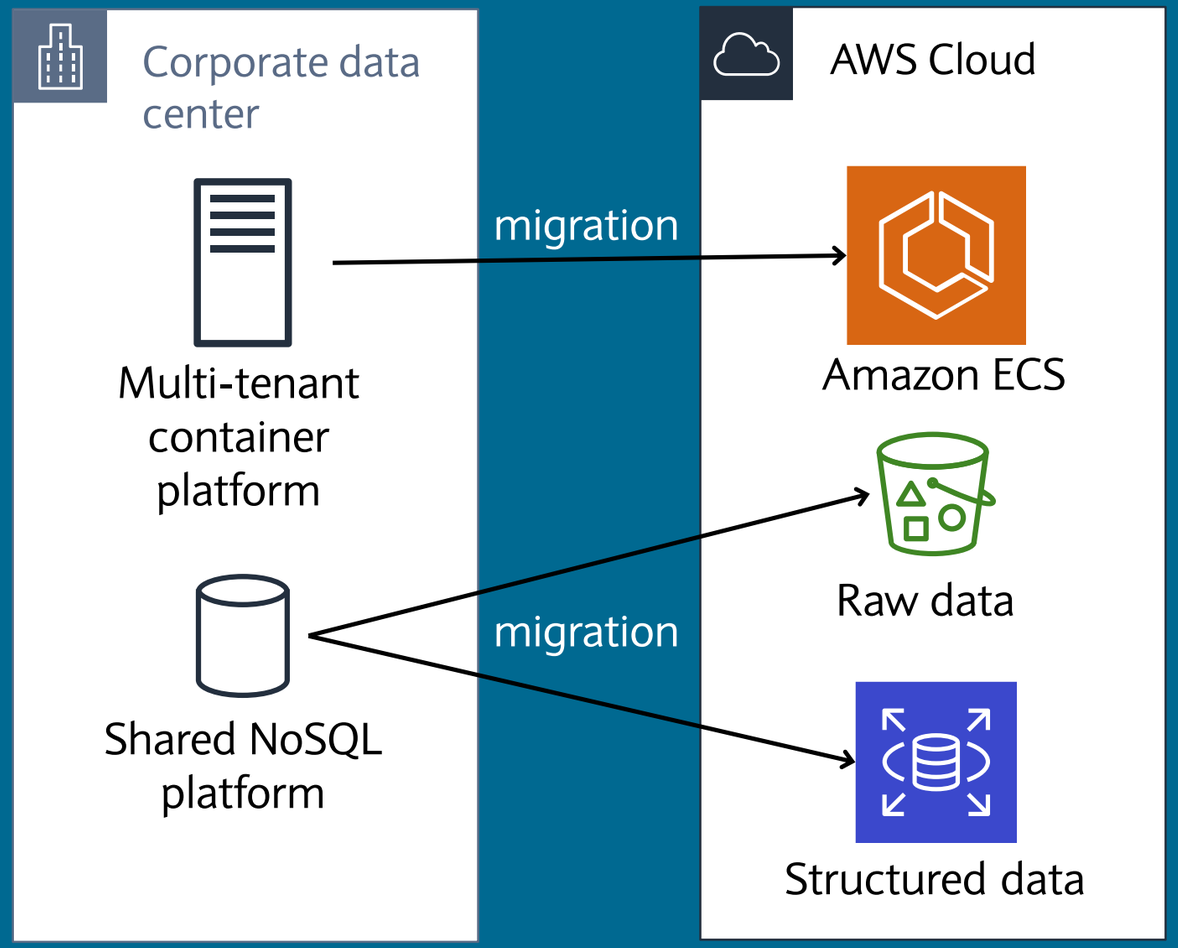
Twitter sentiment analysis developed in Python

Challenge

- Unpredictable load
- Reprocessing times

Benefits of migration

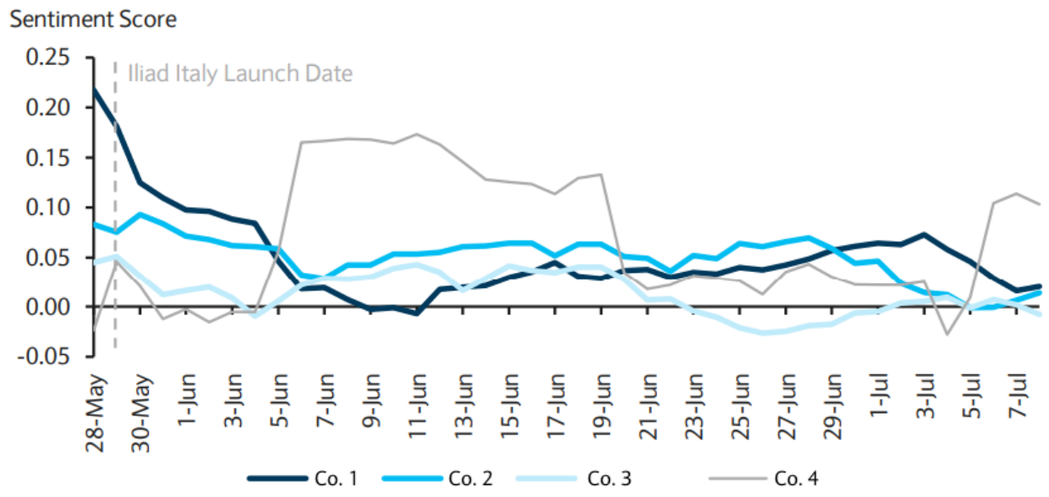
- Elastic compute
- More cost-effective storage



Challenge 4: Sample publication

FIGURE 3

Sentiment of tweets towards wireless carriers in Italy since Iliad's launch



Note: 7-Day Trailing Twitter Sentiment. Source: Twitter, Barclays Research

BARCLAYS

Iliad SA

@IliadItalia sentiment supportive

Our analysis of c.50k tweets since 29 June that suggests that (1) the @IliadItalia mobile launch remains front of mind for consumers almost two months post launch; (2) sentiment towards Iliad in Italy remains positive vs. peers. We see this as supportive of Iliad's ability to maintain notoriety and positive sentiment in online channels, thereby sustaining early launch momentum. We update our forecasts to reflect Iliad's 1m subs milestone in Italy last week, also trimming our French expectations given heightened competitive intensity shows no signs of abating. 2018E revenues are unchanged with Italy upgrades balancing French downgrades, but our EBITDA forecasts fall on higher assumed Italy customers/losses, on higher roaming costs per our bottom-up cost model. We see the biggest opportunity for re-rating as a French operational turnaround and/or inorganic market repair, with >90% of EV in France. 2Q results will likely remain challenged in France, with better opportunities for a rebound in 2H18 per *1H headwinds 2H tailwinds* (05 April 2018). We are OW ILD with our €185 PT (from €195 prior), implying 27% upside potential. ILD ex-Italy trades on 6x 2019E EV/EBITDA, 21x OpFCF and 2% EFCF vs. challenger peers on 7x/14x/6%, respectively.

Summary

Summary of lessons learned

Control & security

- Complex human systems versus complicated problems
- Embed security controls in the application/architecture
- Use of explicit deny can help with security reviews and add resilience

Migration

- Cost optimizations can be found with minimal changes
- Often simpler to migrate all tiers
- Even simple migrations can provide significant benefits
- Can achieve a huge amount with a few AWS services

Future outlook

Increase platform capability

Additional datasets

Greater adoption of managed services

Amazon EMR

Amazon Managed Streaming for Kafka

Amazon Kinesis

Amazon Comprehend

Amazon SageMaker

Amazon Simple Notification Service (Amazon SNS)

Amazon Simple Queue Service (Amazon SQS)

Thank you!

AWS re:Invent

aws

Please complete the session survey in the mobile app.