Track 1|Session 6

建立安全高效的資料分析平台加速金融創新

HC Lo
Solutions Architect
Amazon Web Services

Cliff Chao-kuan Lu
Principal Cloud Architect
EMQ

aws SUMMIT ONLINE

Agenda

The value of data in FSI/Fintech

Building a modern data platform on AWS

The value of data in FSI/Fintech

Data is a strategic asset for every organization

> “The world’s most valuable resource is no longer oil, but data.”

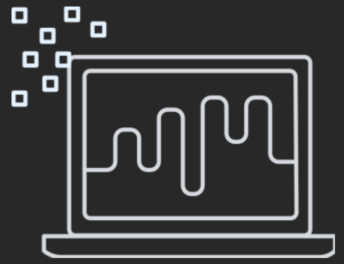
David Parkins, 2017, The Economist

Data creates the next edge for financial institutions

Compliance & Regulatory Reporting

- Data lineage and traceability
- Markets in Financial Instruments Directive (MiFID II)
- Consolidated Audit Trail (CAT)
- Anti-money laundering

Business Analytics



- Gain a holistic view of the business
- Identify market trends and opportunities
- Capture usage data from multiple devices
- Fraud detection

Markets Surveillance & Trading

- Aggregate traditional market data with alternative data
- Markets surveillance
- Portfolio optimization
- Back-testing trading / investment strategies

Customer Experience

- Capture interaction data
- Create targeted products and services
- Provide personalized experiences with timely, tailored messages
- Reach customers via their preferred channel

Core processing → Client facing

Challenge

- Need to handle 150 billion events per day
- Need to run complex surveillance queries over 20+ PB of data to detect and analyze illegal market activity

Solution

- Data lake – S3
- Data processing – Amazon EMR, Amazon Redshift

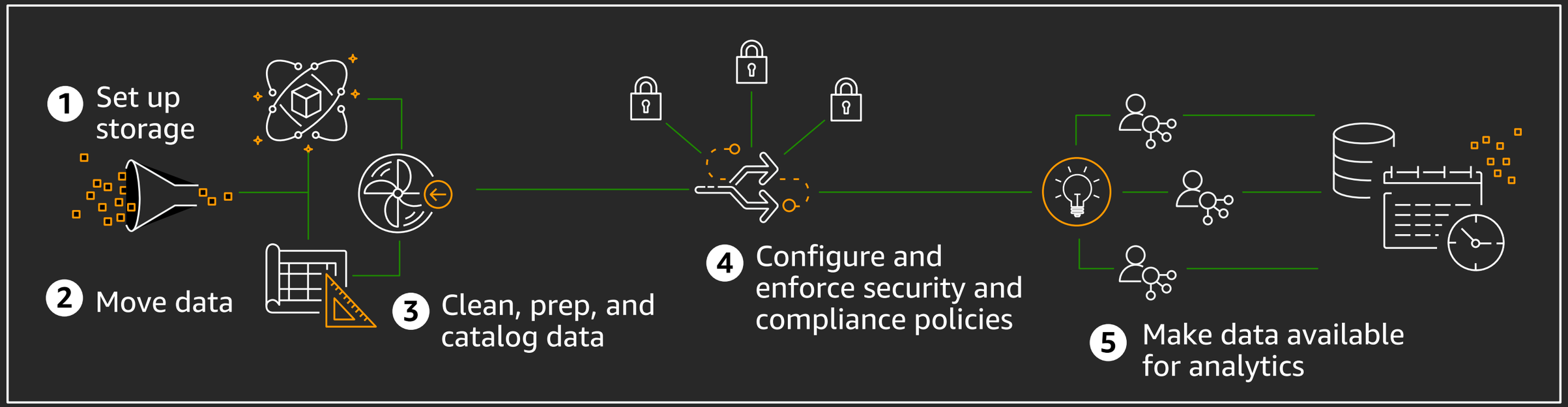
Benefits

- increase agility, speed, and cost savings while also operating at scale
- Estimate to save 10~20 million USD annually

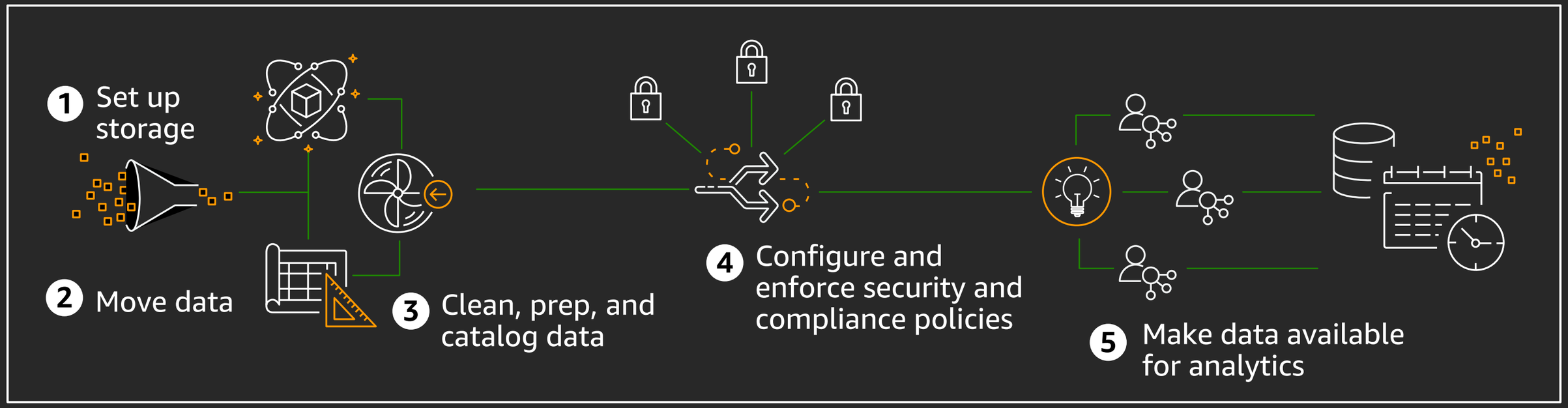
Building a modern data platform on AWS

aws SUMMIT ONLINE

Typical steps of building a data platform



Common considerations

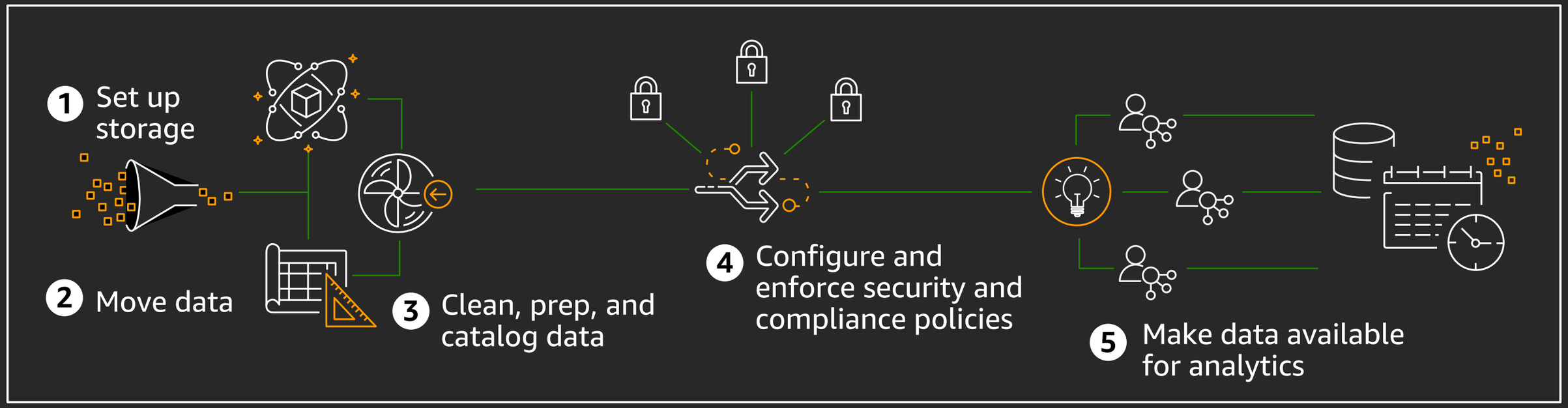


1 Easy to build

2 Comprehensive and open

3 Secure infrastructure

Common considerations



1 Easy to build

2 Comprehensive and open

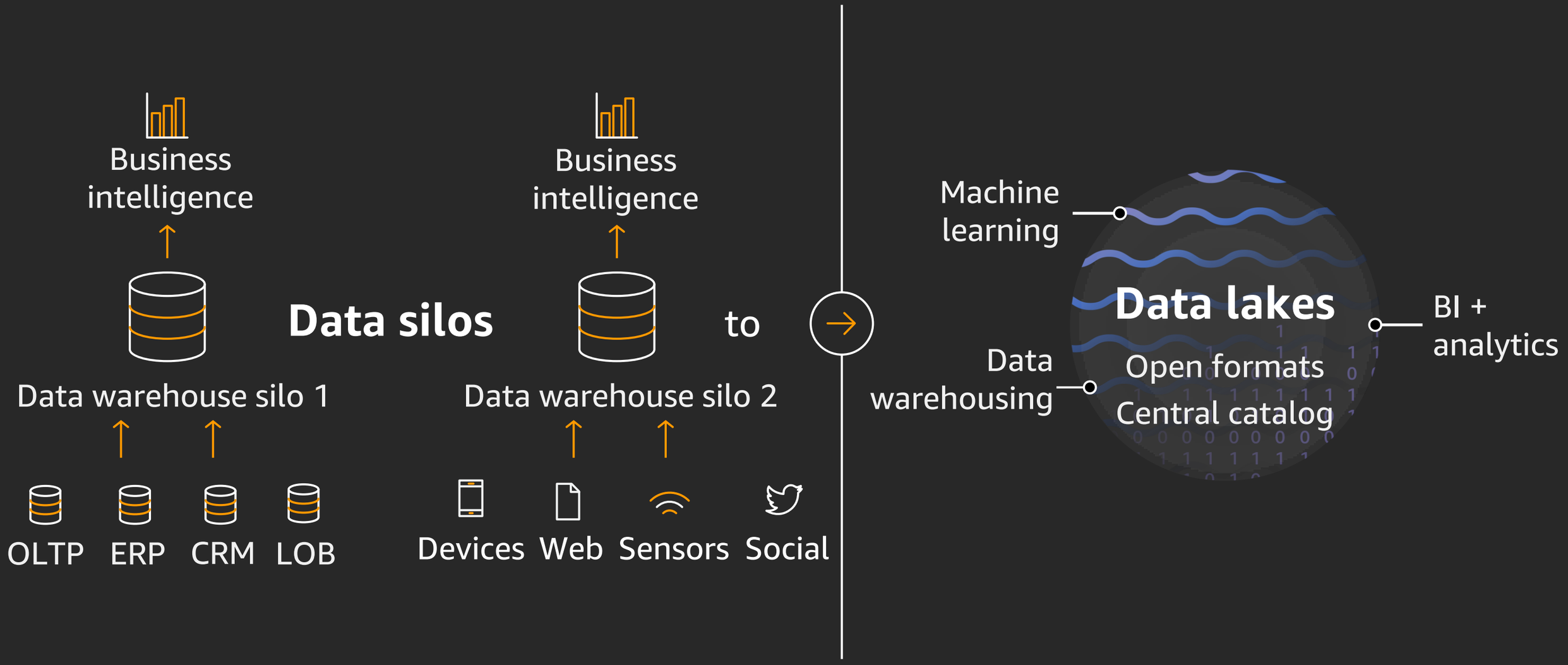
3 Secure infrastructure

Financial institutions are collecting an unprecedented amount of data

- **90% of data worldwide** has been generated in the last five years
- **2.2 million terabytes** of new data is created everyday[1]
- Structured, semi-structured, and unstructured data

1. Axis Corporate, Understanding Big Data in Financial Services, https://axiscorporate.com/us/infographic/understanding-big-data-in-financial-services-infographic/

Traditional data warehousing approaches don’t scale



Build on robust data lake with Amazon S3

- Store any data in any format
- 99.999999999% durability
- Global replication capabilities
- Secure, compliant, and auditable
- Cost-effective storage classes

Choose the right data lake storage class

Raw data	ETL	Production data lake	Online cool data	Historical data
				
Amazon S3 Standard	Amazon S3 Standard	Amazon S3 Intelligent-Tiering	Amazon S3 Standard Infrequent Access	Amazon S3 Glacier or S3 Glacier Deep Archive

Optimize costs for all stages of data lake workflows

Serverless ETL and data integration with AWS Glue

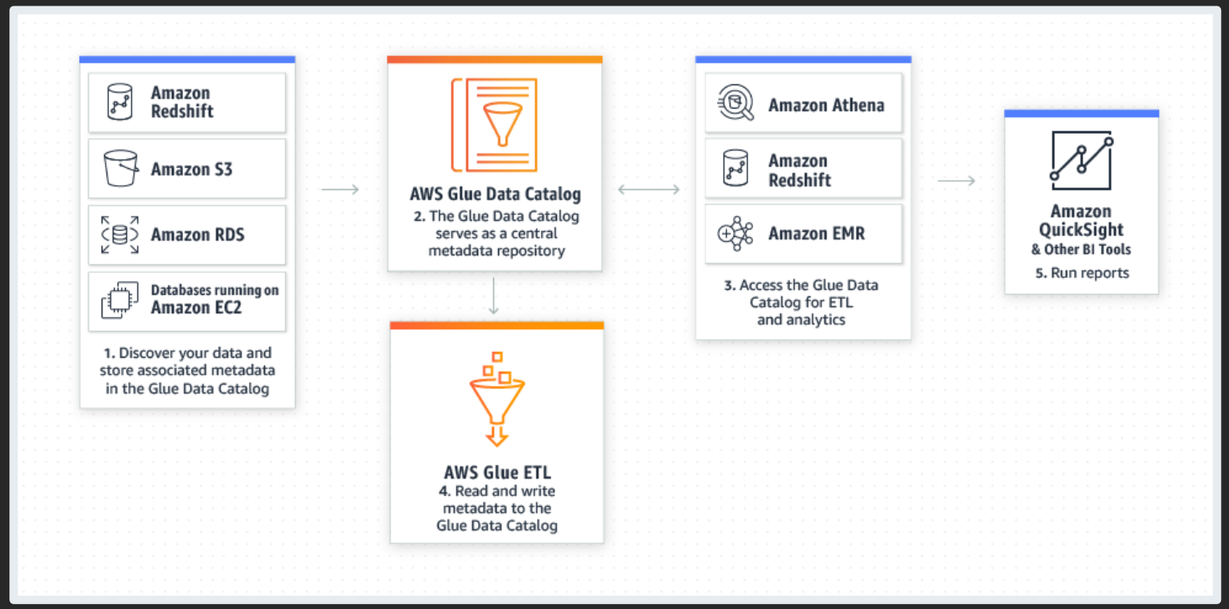
Serverless provisioning, configuration, and scaling to run your ETL jobs

Pay only for the resources used for jobs

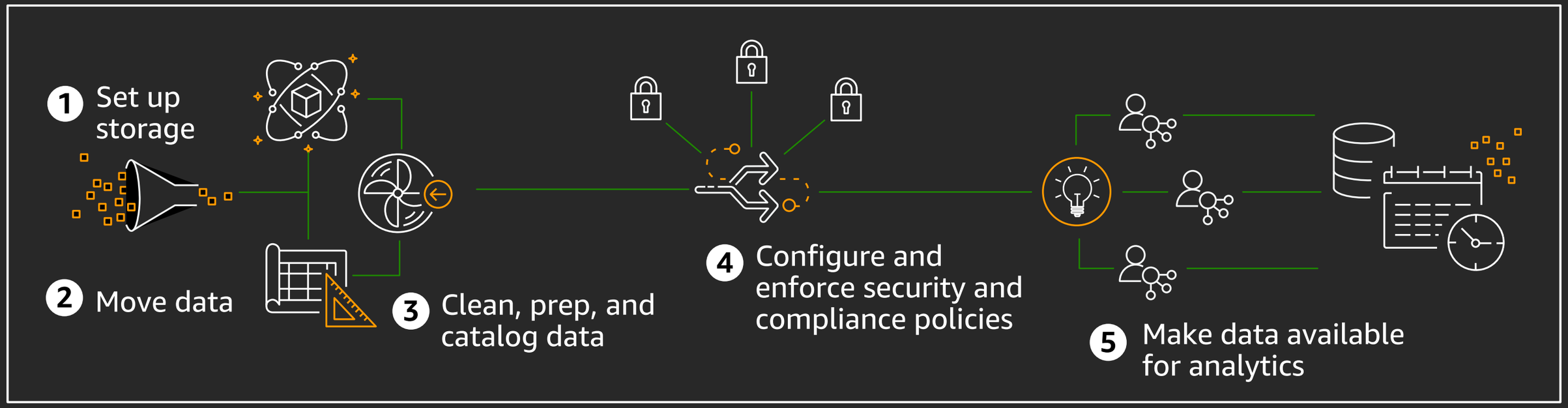
Automates the effort in building, maintaining, and running ETL jobs

Makes it easy to schedule recurring ETL jobs, chain multiple jobs together, or invoke jobs on-demand

Manage the dependencies between your jobs, automatically scales underlying resources, and retries jobs if they fail



Common considerations



1 Easy to build

2 Comprehensive and open

3 Secure infrastructure

Comprehensive data analytics services

Real-time analytics

Operational analytics

Big data processing

Data warehousing

Interactive query

Comprehensive data analytics services

Real-time analytics

Amazon Kinesis Data Analytics

Operational analytics

Amazon Elasticsearch Service

Big data processing

Amazon EMR

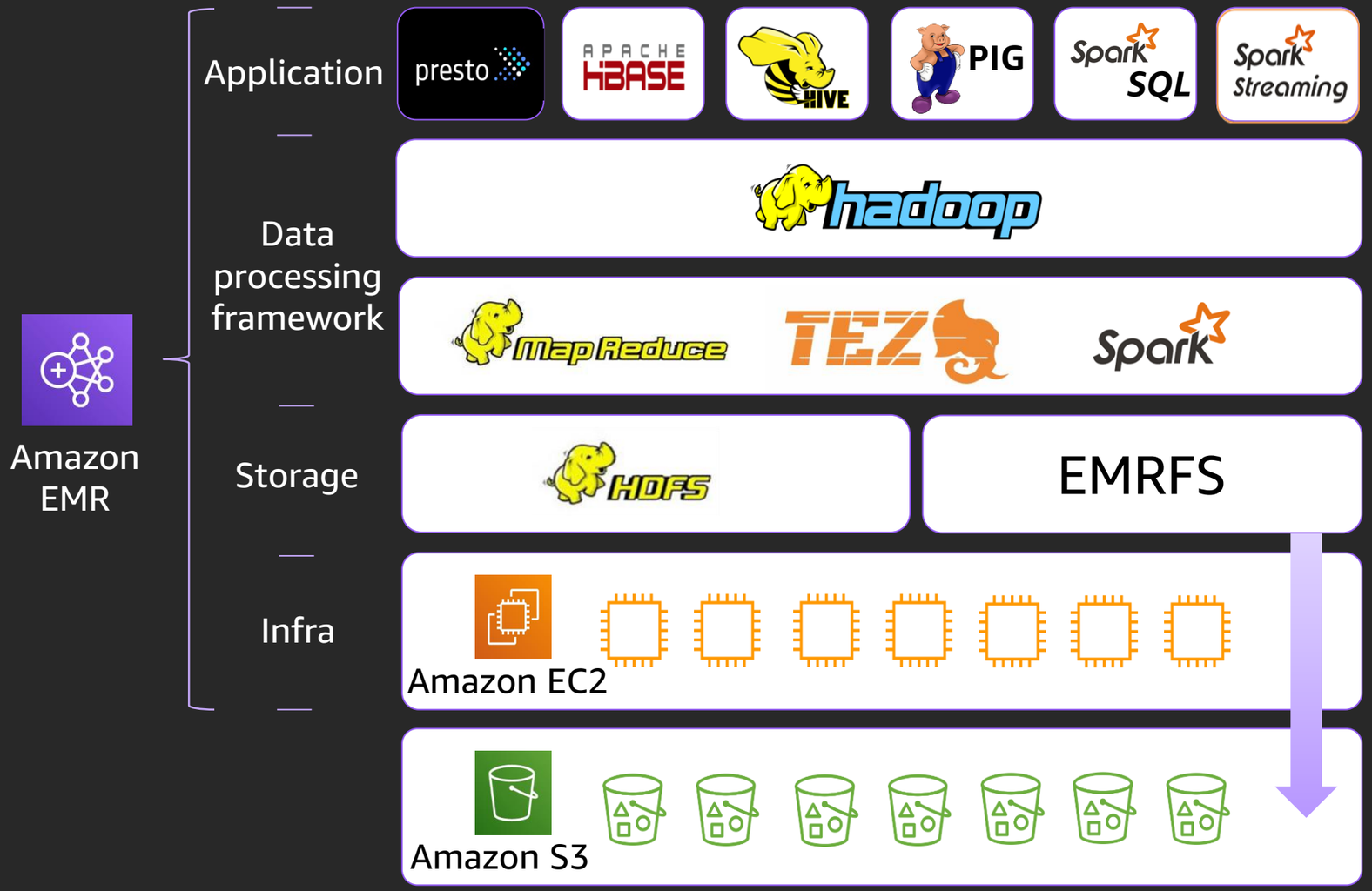
Data warehousing

Amazon Redshift

Interactive query

Amazon Athena

Big data processing: Amazon EMR



Managed petabytes of unstructured or semi-structured data

Support 21 open source projects, including Apache Hadoop, Spark, HBase, Presto, etc.

Decouple compute and storage

Enterprise-level security and reliability

Cost optimization

Data warehousing: Amazon Redshift

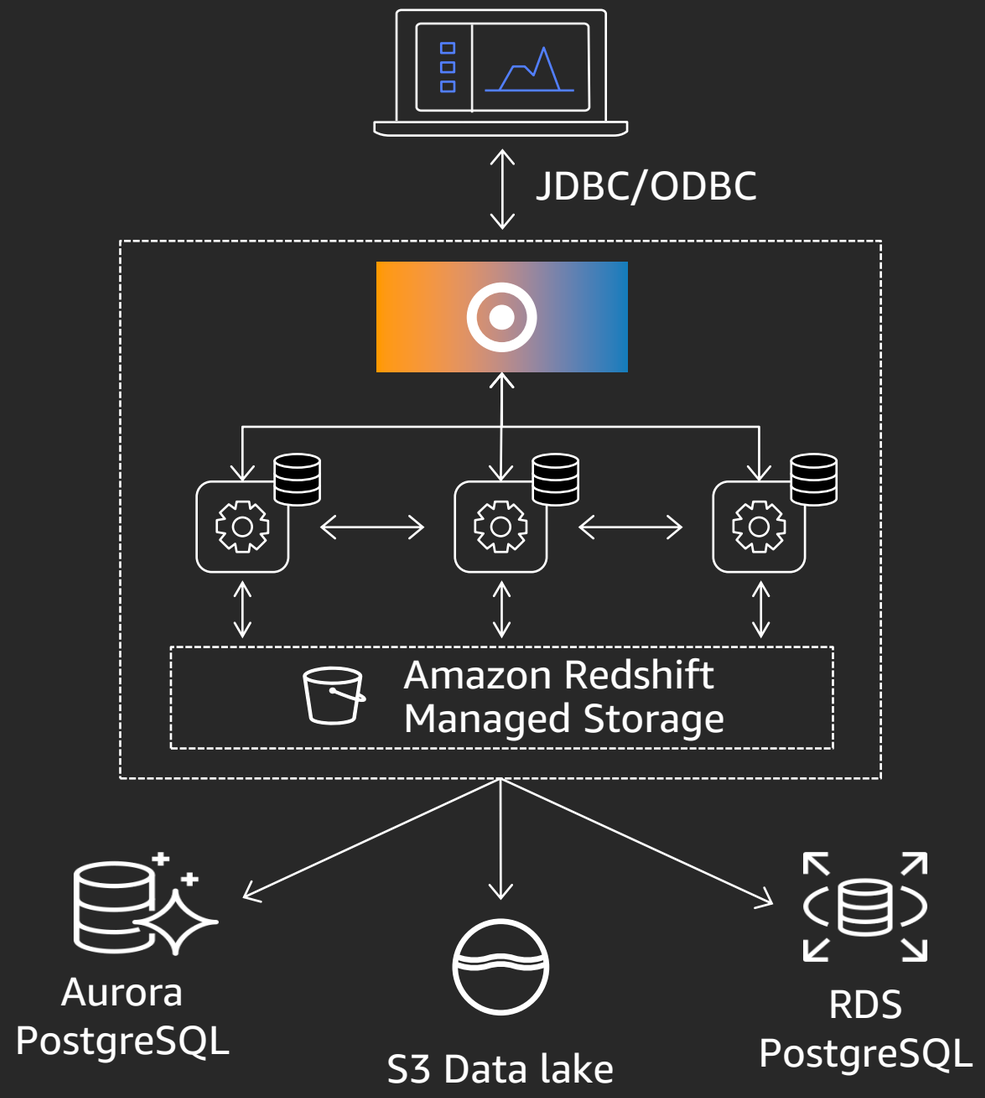
Analyze exabytes of data across data warehouses, data lakes, and operational databases

Massively parallel data processing

Columnar storage, data compression, and zone maps reduce the amount of I/O needed to perform queries

Cost-optimize workloads by paying compute and storage separately

Security out of the box, at no extra cost



Interactive query: Amazon Athena

Query instantly

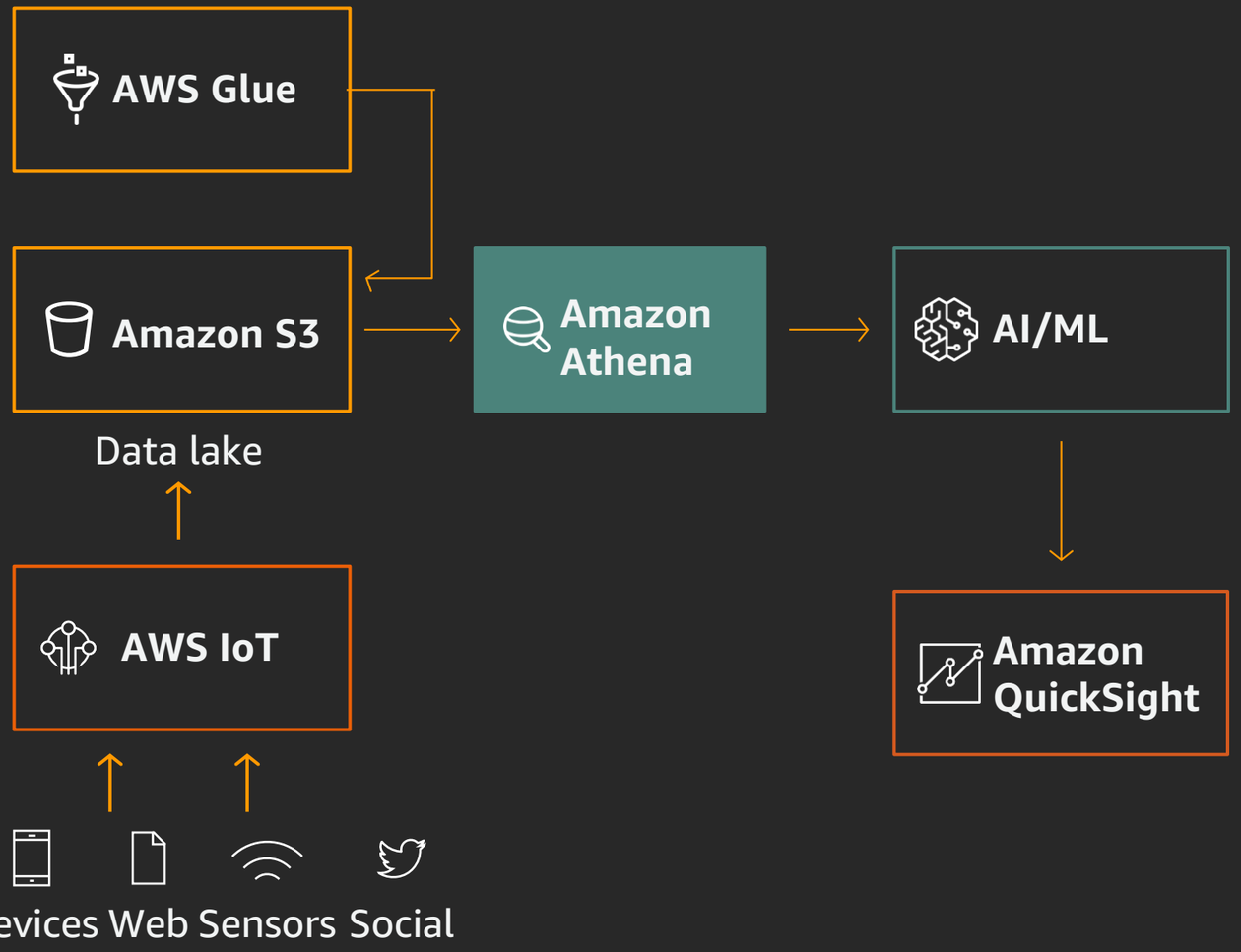
Point to S3 and start querying with standard SQL queries

Pay per query

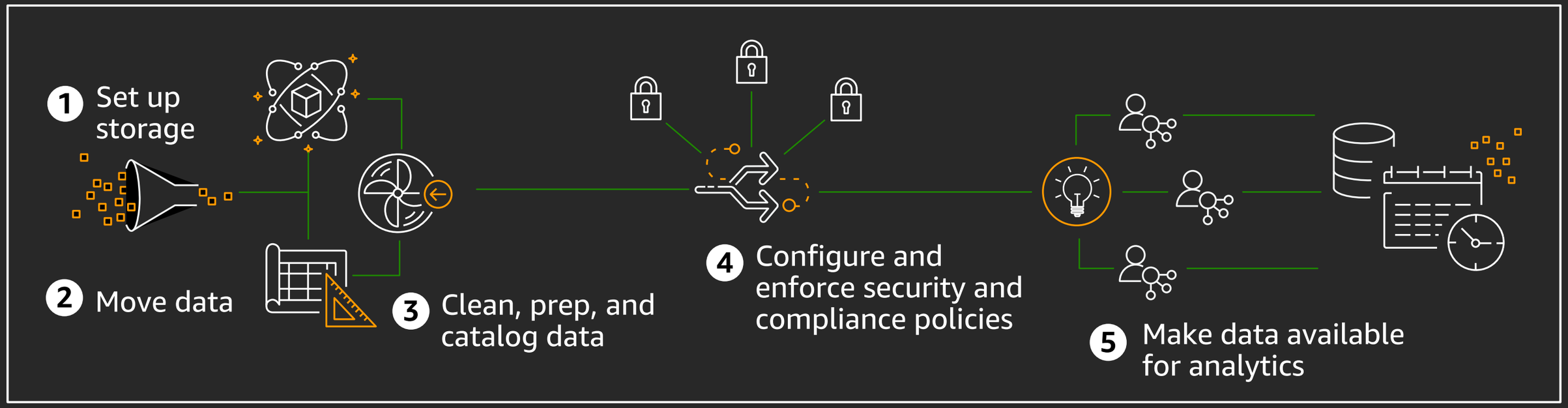
Save 30%–90% on per-query costs through compression

Serverless, zero infrastructure, zero administration

Scales automatically based on complexity of queries



Common considerations



1 Easy to build

2 Comprehensive and open

3 Secure infrastructure

Common security considerations for data lakes

Compliance programs

- What regulatory and data privacy issues should we consider when building a data lake?
- What services should we use to meet our compliance obligations?

Multi-account support

- Can our data lake scale to support a multi-account strategy?
- Can data stewards have ownership of their data and metadata once it's in the data lake?

Authentication

- Can our users access the data lake in accordance with corporate authentication standards?
- Can data lake user interfaces integrate with our Active Directory?

Authorization

- Does the data lake support role-based authorization for accessing data and metadata in the data lake?
- How do we ensure that users only access data that they are allowed to see?

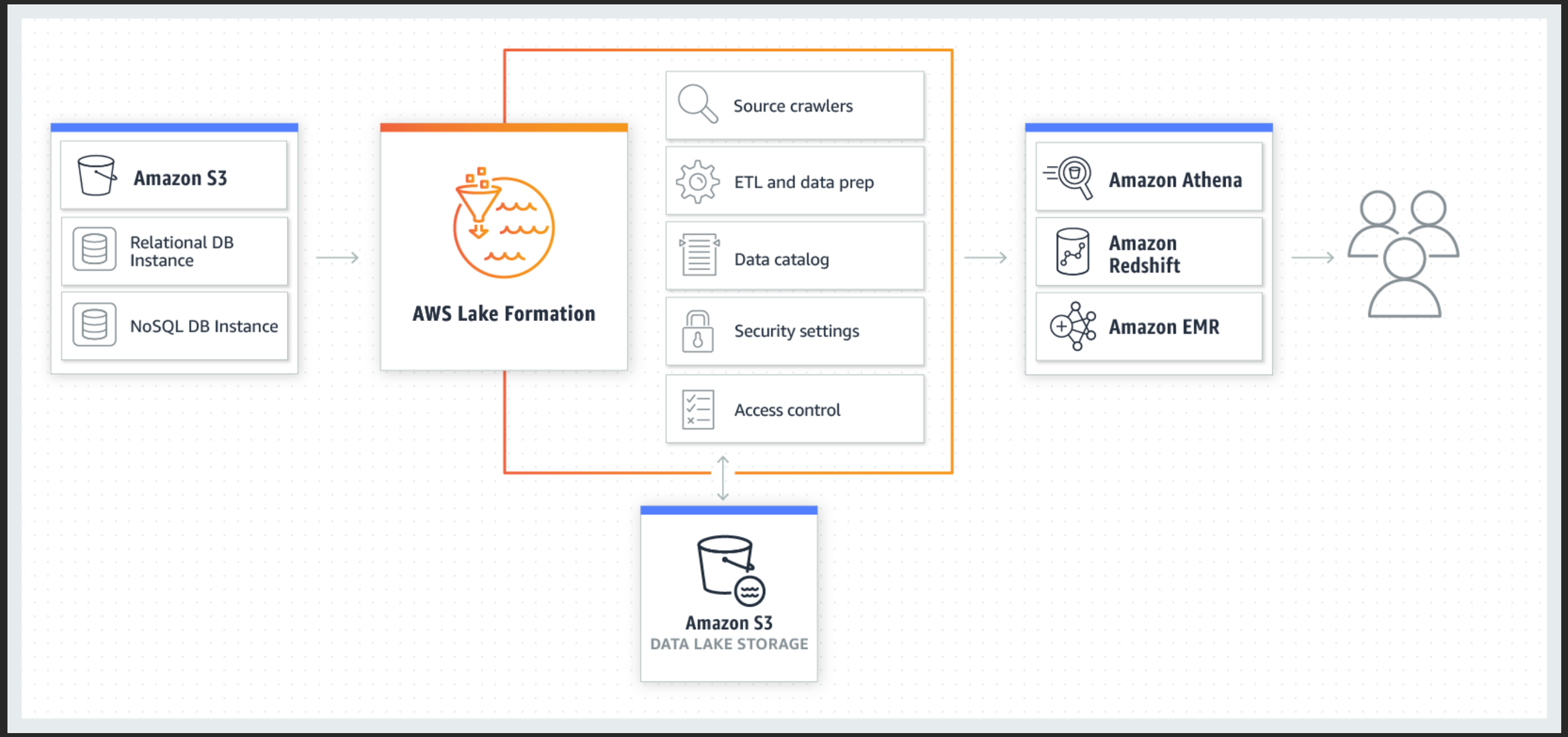
Encryption

- Does the data lake architecture support our corporate encryption standards for data at rest and in motion?
- Does the data lake integrate with our key management service?

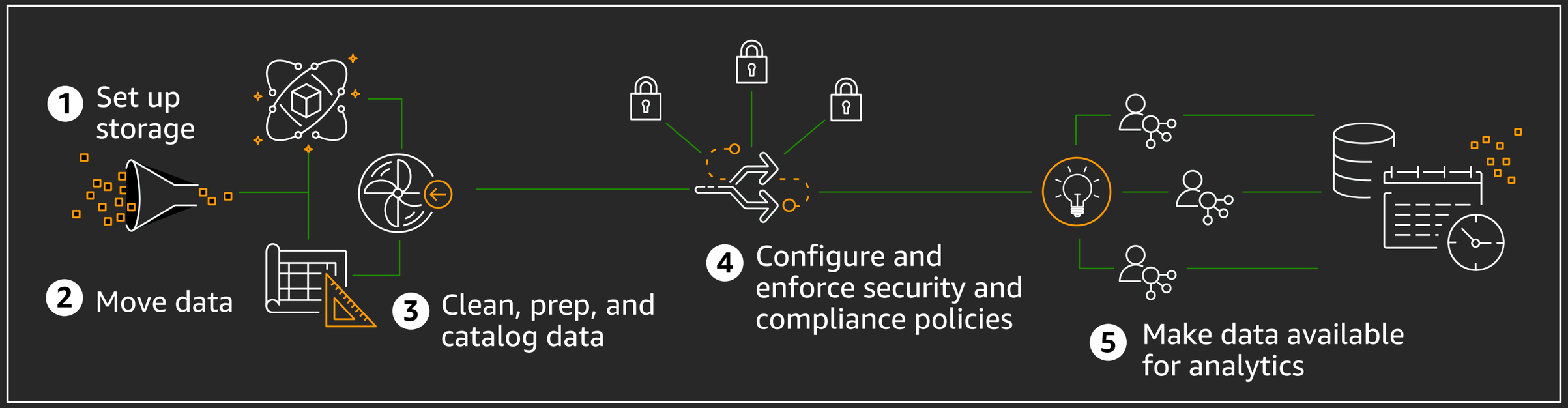
Private network connectivity

- Can all traffic to and from the data lake transfer over private and secure network links?
- Can we block all internet access to and from our data lake?

AWS Lake Formation: simplify security management



Summary: turn data into insights



1 Easy to build

2 Comprehensive and open

3 Secure infrastructure

Let’s welcome Cliff Lu, Principal Cloud Architect of EMQ to share their journey on AWS!

aws SUMMIT ONLINE

Agenda

EMQ 公司介紹

原分析流程

引入無伺服器資料分析服務

總結

EMQ 公司介紹

EMQ

A cross-border settlement solution for any financial transaction

FinTech since **2014**

Licensed in **Hong Kong**, **Indonesia**, **Singapore**, and **Taiwan**

15+ bank partners

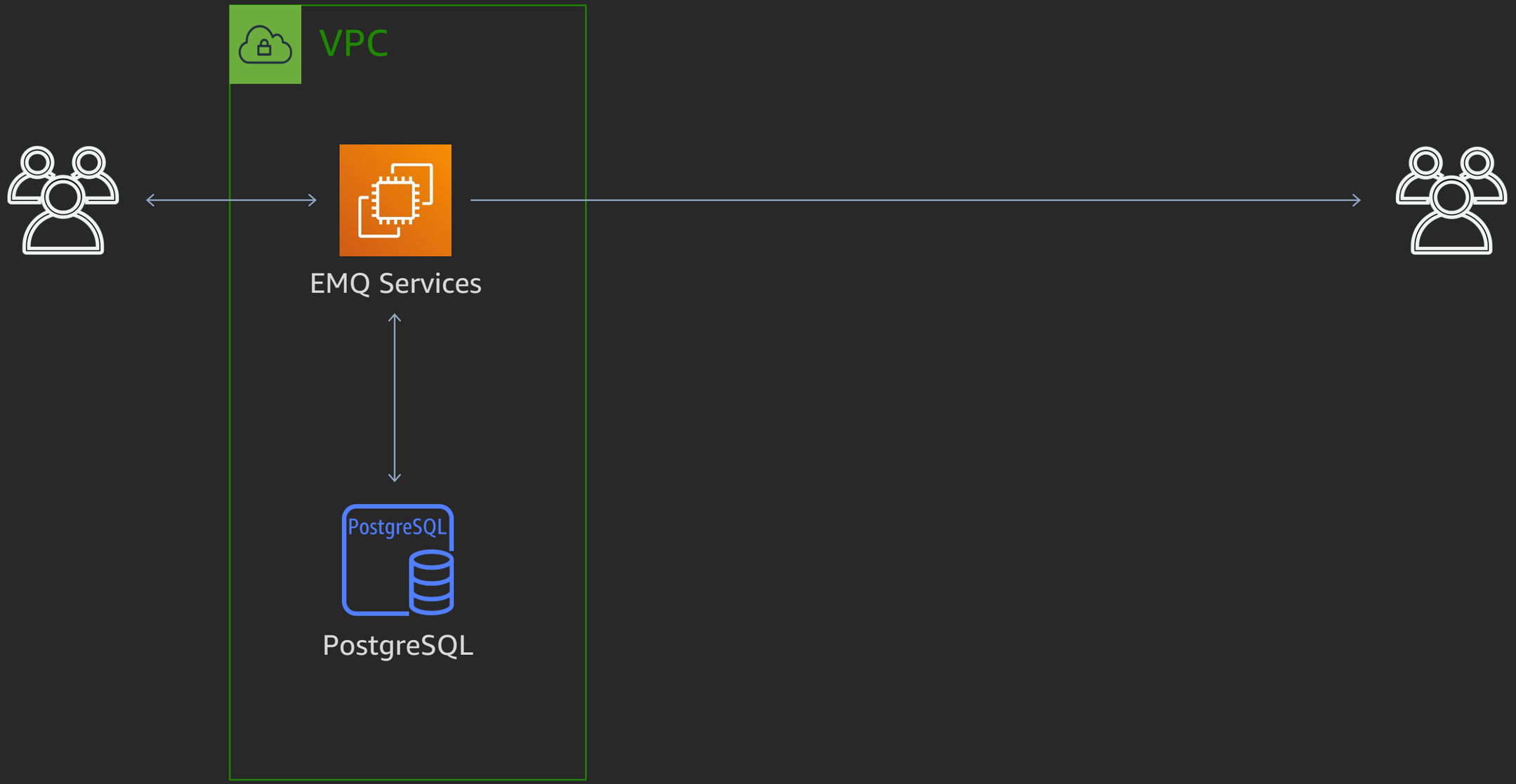
50+ payout countries

15,000+ cash pickup location

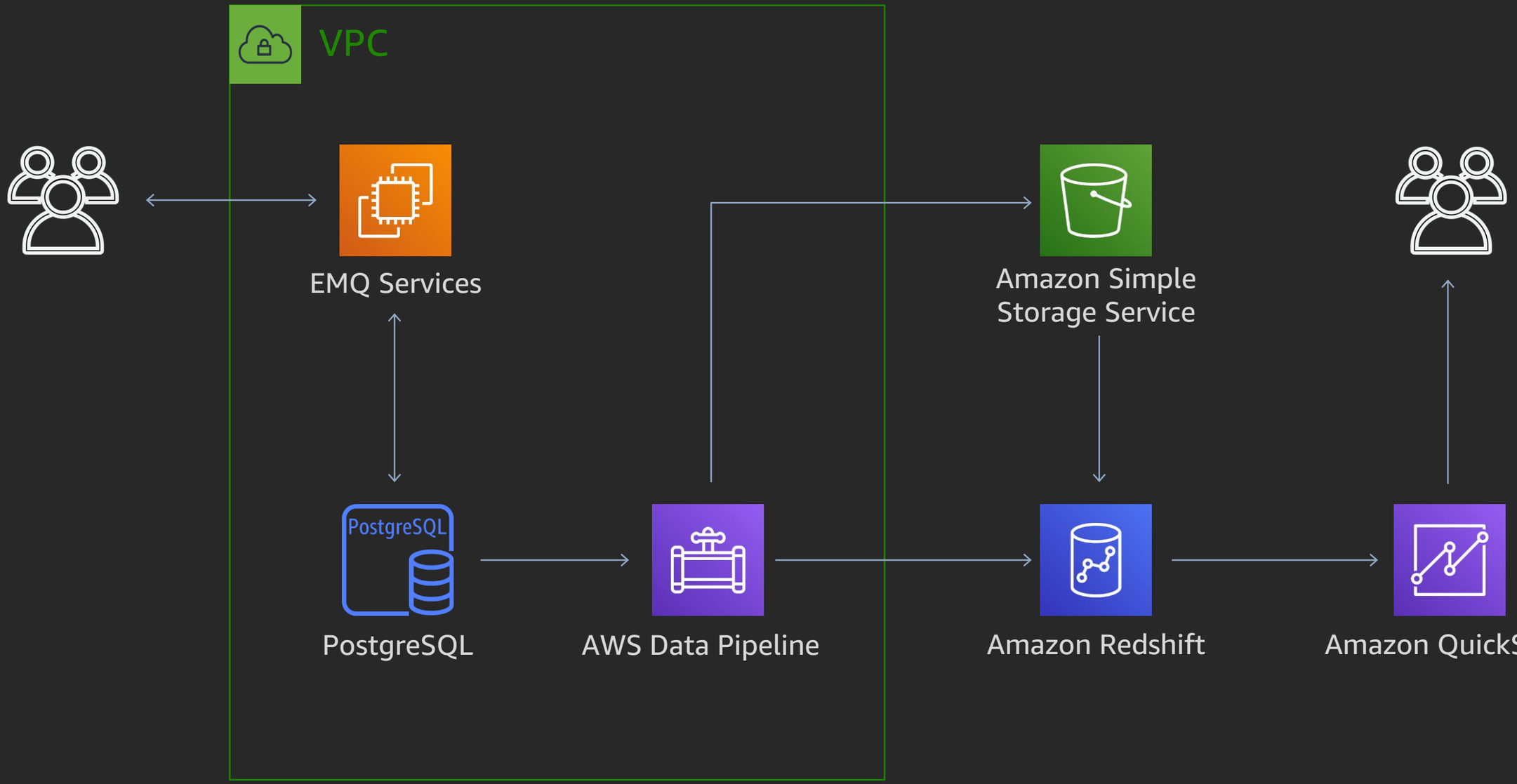
200,000+ bank branches

原分析流程

EMQ 由資料驅動業務



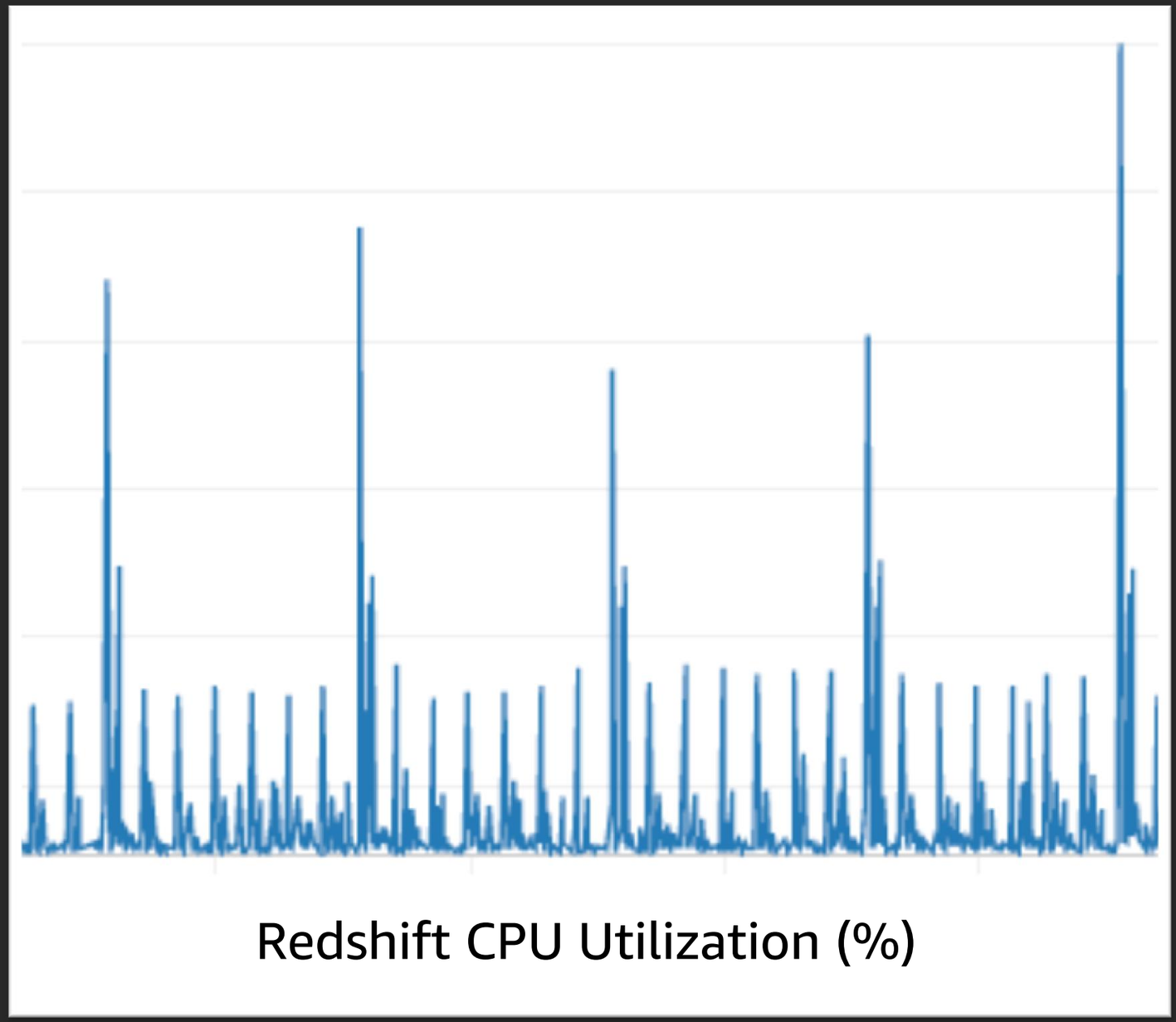
運用 Amazon Redshift 後查詢效率提升



使用 Data Pipeline 與 Redshift

- 需要 S3 作為中介
- 由 Data Pipeline 配置 Redshift 的結構
- Redshift 本身可用於去除重複資料列
- 在運行報告的時間外，叢集會閒置；因此平均利用率低
- T 系列有助於控制成本

尋找更適合現階段的資料分析架構



技術選型目標

- 支持多種輸入、輸出介面；
  可由PostgreSQL 輸入、由 QuickSight 查詢
- 使用主流框架，或支持以主流語言操作資料
- 日常維運門檻必須低
- 資訊安全需容易稽核、配置
- 效能門檻不高，但需要...
  - Output Partitioning
  - 差異備份
  - 資料操作，Join / Flatten / Remove / Mask

Output Partitioning

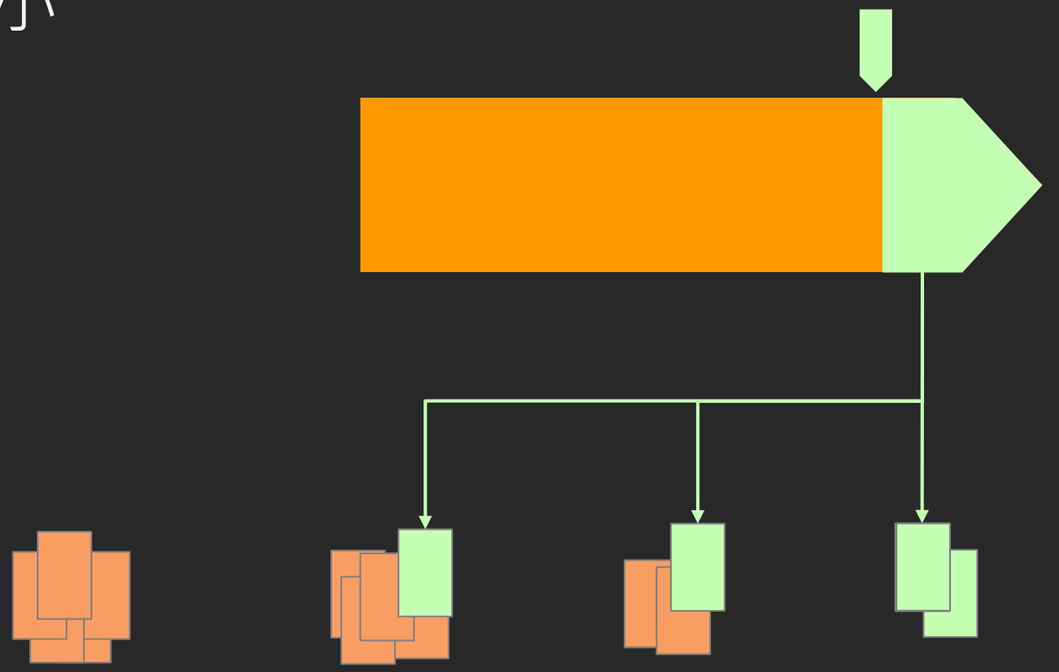
查詢時帶入過濾條件，可減少載入的資料量，加速查詢並減低成本

Column name	Data type	Key
id	string	
type	string	
actor	struct	
repo	struct	
payload	struct	
public	boolean	
created_at	string	
org	struct	
year	string	Partition (0)
month	string	Partition (1)
day	string	Partition (2)

https://aws.amazon.com/tw/blogs/big-data/work-with-partitioned-data-in-aws-glue/

差異備份

- 只讀取更新的資料列，載入資料量小
- 要求遞增欄位
- 大多 **ETL** 工具包含類似機制
- 部分服務原生支持資料合併

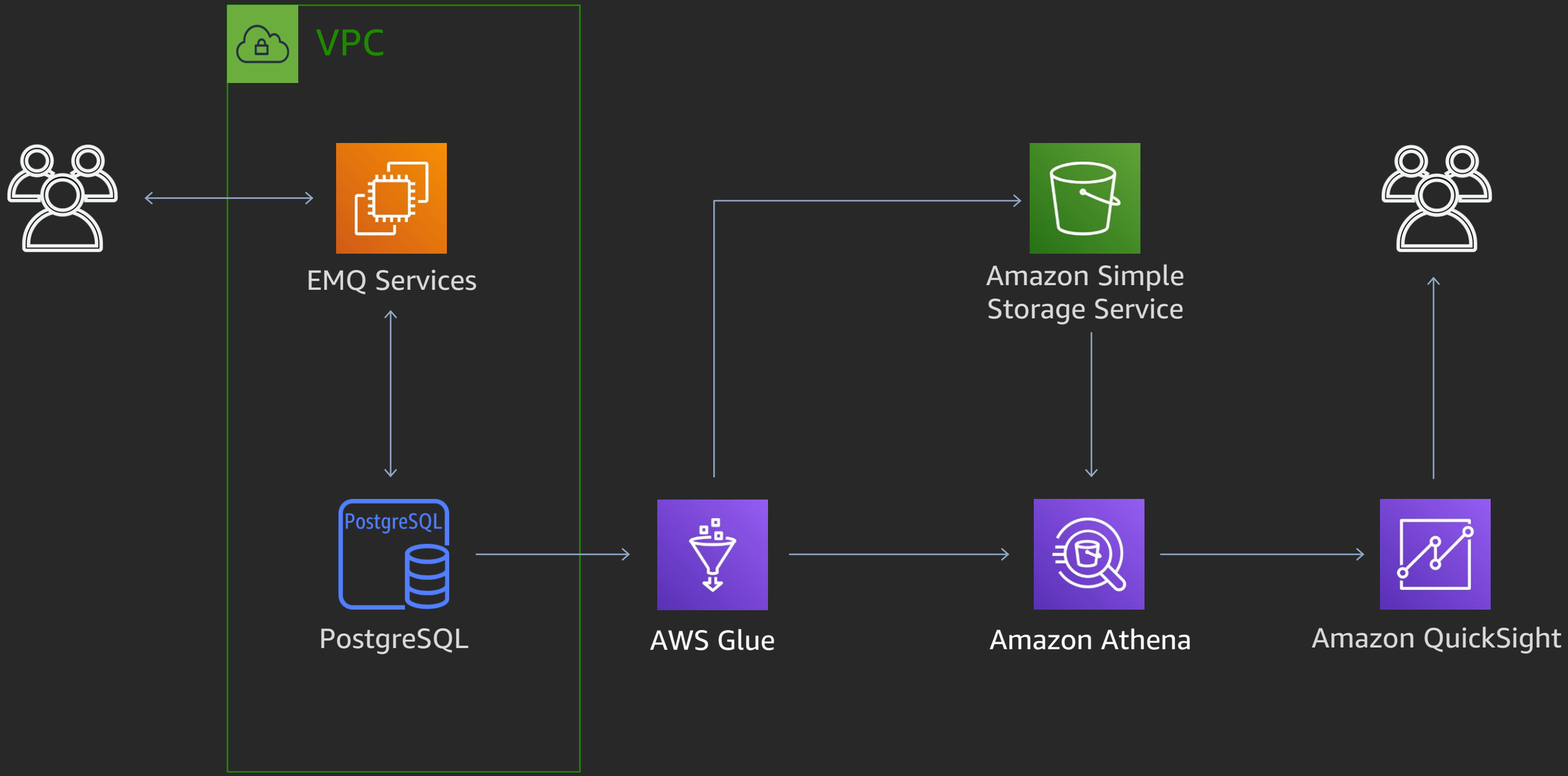


候選名單

	Data Pipeline	Database Migration Service		Glue
儲存 / 分析服務	Redshift	Redshift	S3 / Athena	S3 / Athena
Output Partitioning	不支持但夠快	不支持但夠快	不支持	支持
差異備份	任意遞增欄位	需要 Primary Key		~~遞增 Primary Key~~ 任意遞增欄位
資料操作	Shell Script	不支持 JSON 等結構操作		Python / Spark

引入無伺服器資料分析服務

EMQ 由資料驅動業務: 新架構、無伺服器



Why AWS Glue

- 無伺服器，託管服務
  - 計費模型單純：用多少，付多少
- 支持多種資料來源及輸出格式
  - Glue 的輸出可以再成為 Glue 的輸入
- Glue 妥善整合其他 AWS 服務
  - 容易學習、理解
  - 稽核與配置檢查相對容易
- 抽象結構容易理解、調度功能符合需求
  - Job 可運行 Python 或 Spark

Pushdown Predicate

- Spark SQL 在資料源過濾查詢條件
- 若透過索引完成，可縮短執行時間，並減少運算、傳輸量

- 目前還不支援 ODBC / JDBC 資料源

Why Amazon Athena

- 無伺服器，託管服務
  - 計費模型單純：用多少，付多少
- 透過 SQL 介面查詢資料
- 與其他 AWS 服務整合密切
  - 可用於創建新資料集 (S3) 與資料表 (Glue)
  - 查詢結果儲存於 S3，容易管理與重用
  - 透過 QuickSight 查詢並創建報表

In The Works – 持續優化費用與效能

- 我們的查詢仍有待優化，目前有部分 QuickSight 操作超時
  計劃引入 Partition、暫存表、並將複雜的 Join 移至 Glue 以改善
- 若無法優雅解決差異備份問題，將引入或轉換至 EMR
  - Pushdown Predicate for JDBC / ODBC
- 費用與效能
  - 大多業務報告以 Glue / Athena 運行，持續向 best practices 靠攏
  - 正與分析師協同重構，預估可提供每小時資料更新，並將費用減低爲 1/8

總結

總結 – EMQ 引入無伺服器資料分析服務的效益

- 減緩學習曲線
  - 預設功能堪用、日常維運門檻低
  - 整合其他 AWS 服務，能承襲已有的知識
  - 底層為開源專案，兼作爲學習的基石
- 快速迭代，持續進化
  - 計費模式讓 EMQ 不需在 Day1 預估用量，可以嘗試各種架構設計
  - 在 AWS 生態系內外有多種服務可整合、協作、相互替代
  - AWS 提供的新服務與新功能，讓 EMQ 的選項益加豐富

Thank you!

HC Lo
hclo@amazon.com

Cliff Chao-kuan Lu
clifflu@emq.com

aws SUMMIT ONLINE