AWS Data Governance Demystified

[Network, Security,Privacy &

Data Access Management]

AWS Big Data Demystified #4

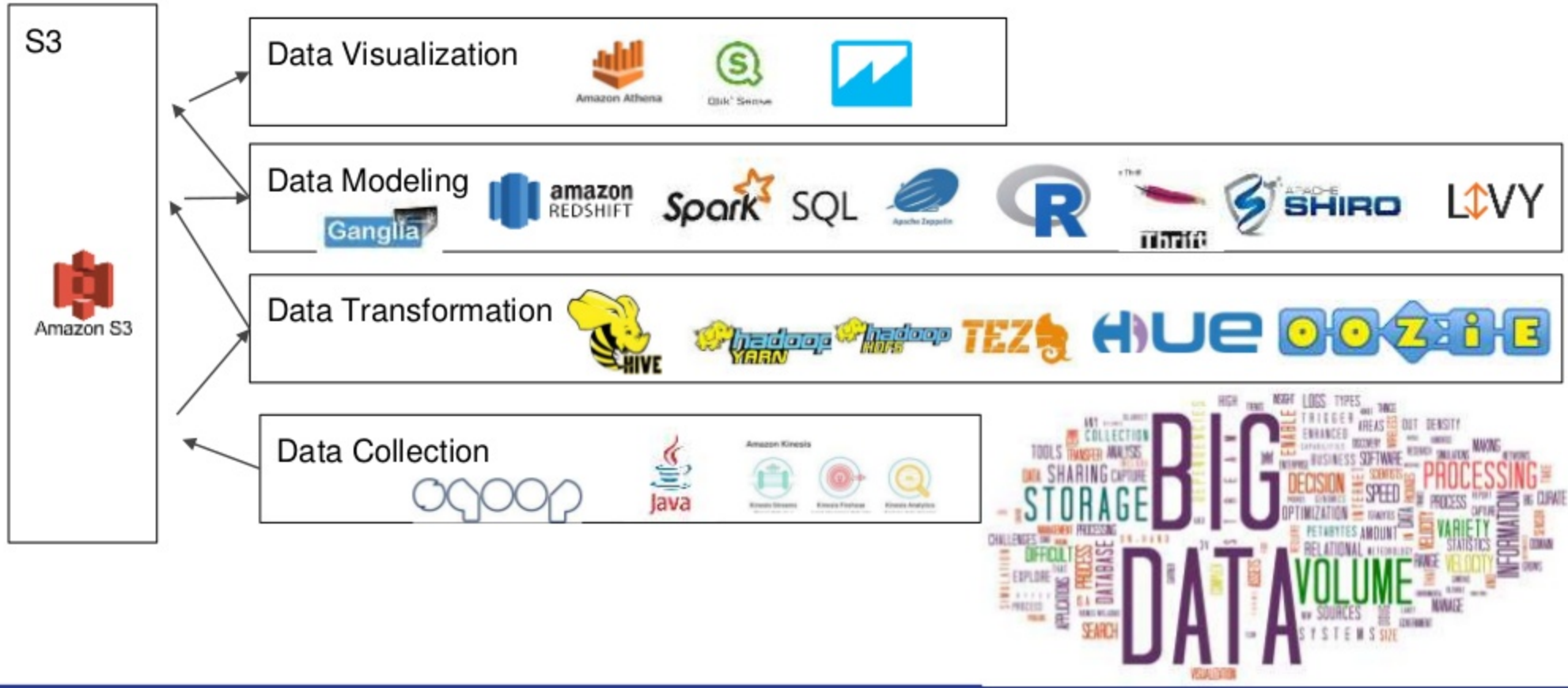
Omid Vahdaty, Big Data Ninja

Agenda

- Disclaimer
  - I am not a network and security expert
  - This is not a security and network lecture
- Agenda….
  - All the possible consideration for my data? (access,regulation, availability etc) == Data Governance
  - What are the architecture implications?
- My Advice:
  - Stick to high level overview
  - Remember the topic not the details.
  - ASK me questions during the lecture!

TODAY'S BIG DATA APPLICATION STACK
PaaS and DC...

Big Data Generic Architecture | Summary



Before I forget: for new AWS accounts....

- Disable unused regions via IAM
- Set the the limit for the instances your are using , e.g 50 instances
- Set the limit for the instances your are not using to 0!
- Remove access key secret key for root account
- Don't use root account
- Use MFA

Data Governance…

Data Security Level

Data governance is the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include **availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise** such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be **used by the entire organization.**

Big Data Security: Nobody likes it… But…

AWS Regulation and more...

- https://aws.amazon.com/compliance/programs/
- PII
- GDPR

AWS Security options in general

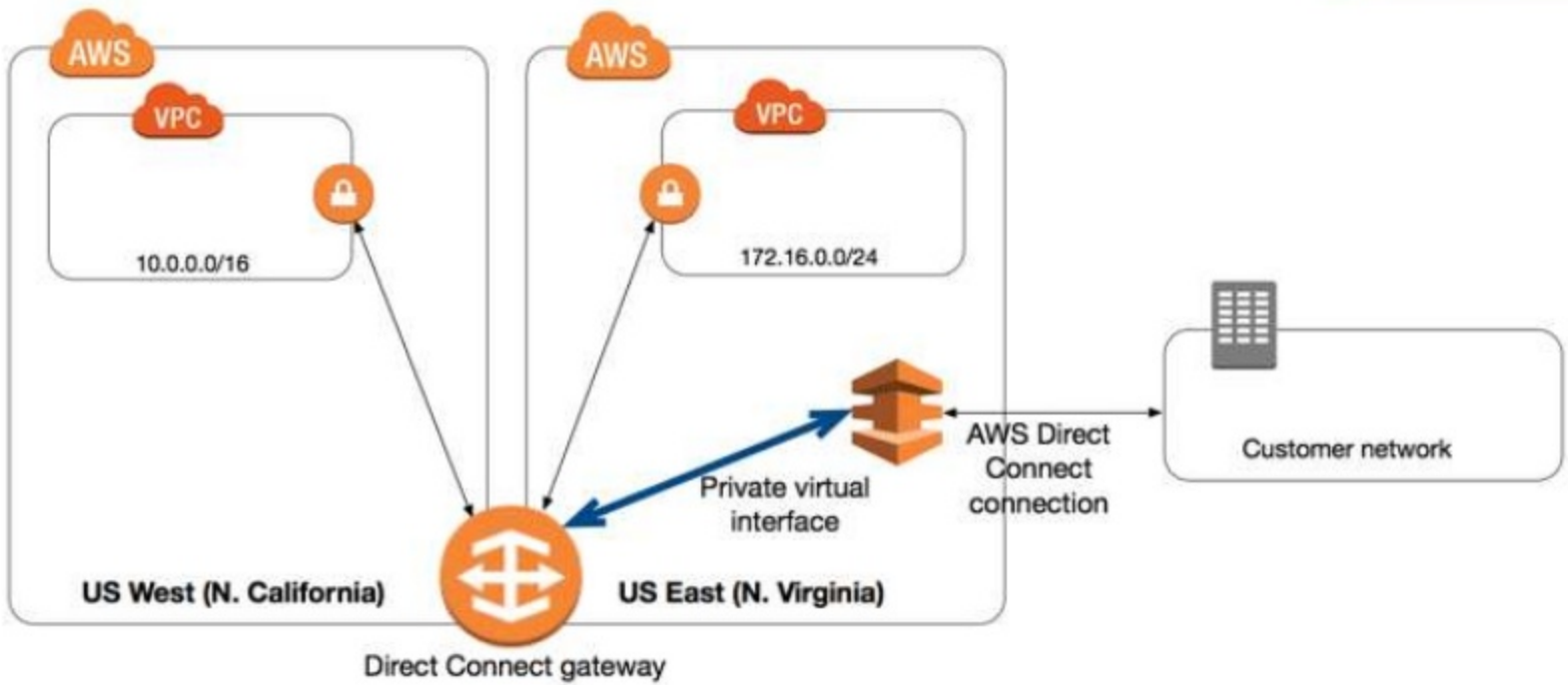
[sorry, not covering everything…]

- IAM : Identity management
  - IAM : Identity management
    - User, policy, roles,group, least privileges, MFA
  - Key Management Service
    - Server Side
    - Client side
  - Disable Data centers, unused instance families,
  - Limit resources
  - Account segregation
  - **Identity based policies!**
- Cloud trail: enables governance, compliance, auditing, and risk auditing
- S3: Resource management (e.g s3)
  - Write only, read only, no delete
  - Versioning, encryption,life cycle policy

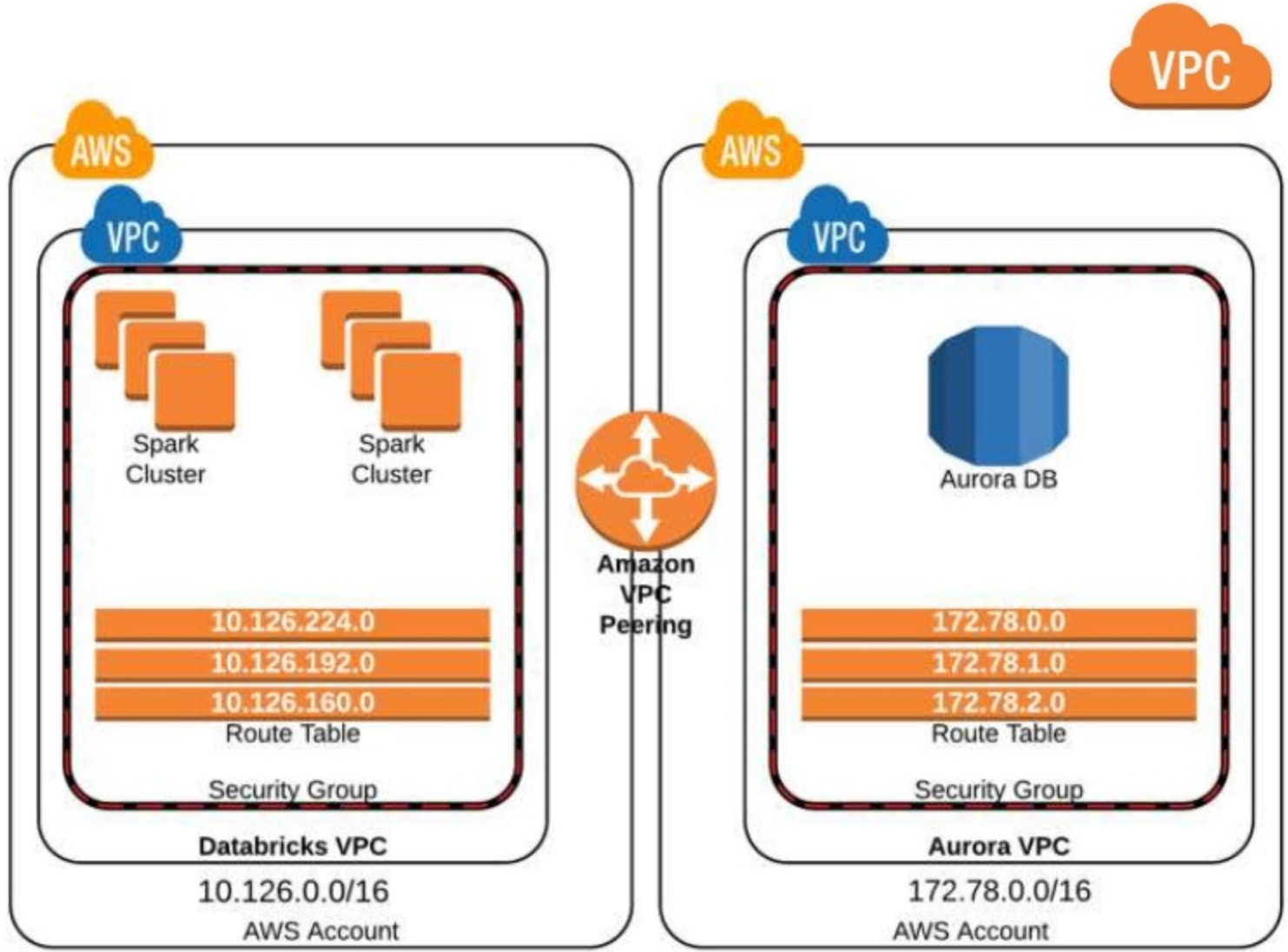
AWS Network options in general

- VPC
  - Create a Non Default VPC
  - Private network + Bastion host + ALB
  - Public subnet vs private subnet
  - VPC Endpoints
  - NACL
  - SG
  - IG
  - VPC peering
  - Site 2 Site - it is per VPC…. choose carefully
- Direct Connect VS HTTPS
- Cloudfront with GEOlocation protection
- https://amazon-aws-big-data-demystified.ninja/2018/06/27/aws-enterprise-grade-networking-security-what-are-your-options-to-protect-your-bigdata/

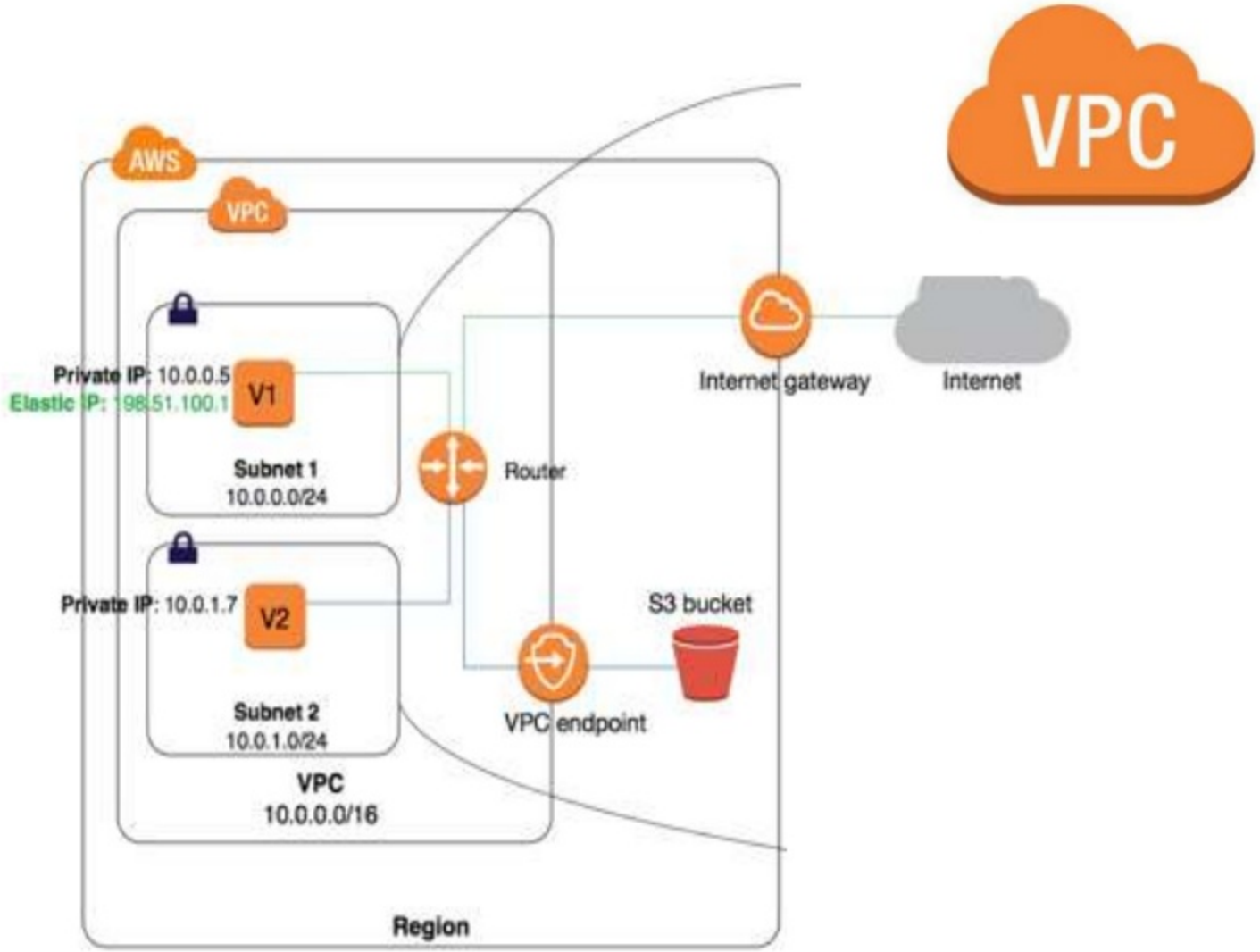
AWS Direct Connect



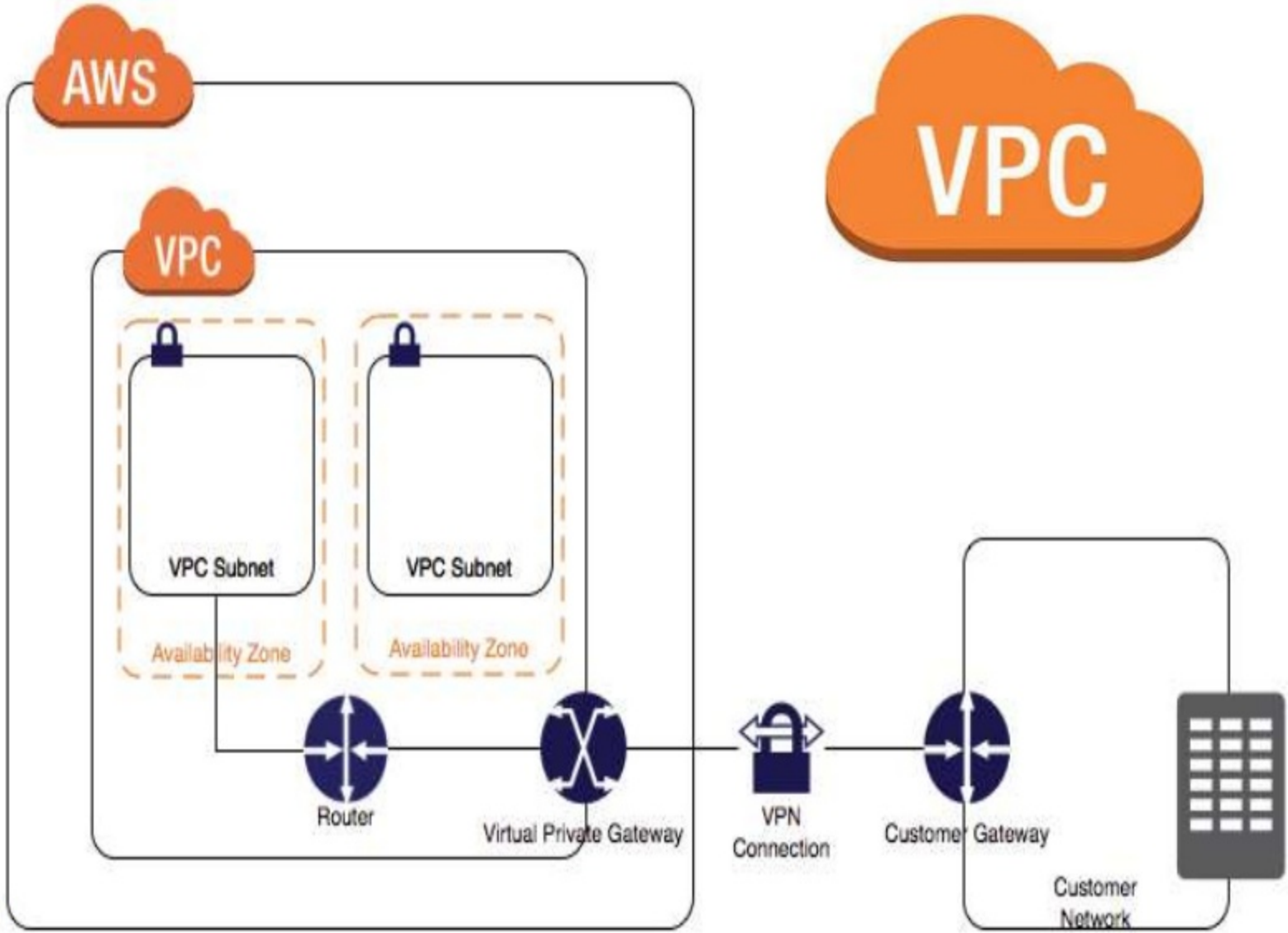
VPC peering


VPC
AWS
VPC
Spark Cluster
Spark Cluster
10.126.224.0
10.126.192.0
10.126.160.0
Route Table
Security Group
Databricks VPC
10.126.0.0/16
AWS Account
Amazon VPC Peering
AWS
VPC
Aurora DB
172.78.0.0
172.78.1.0
172.78.2.0
Route Table
Security Group
Aurora VPC
172.78.0.0/16
AWS Account

S3 Endpoint Example

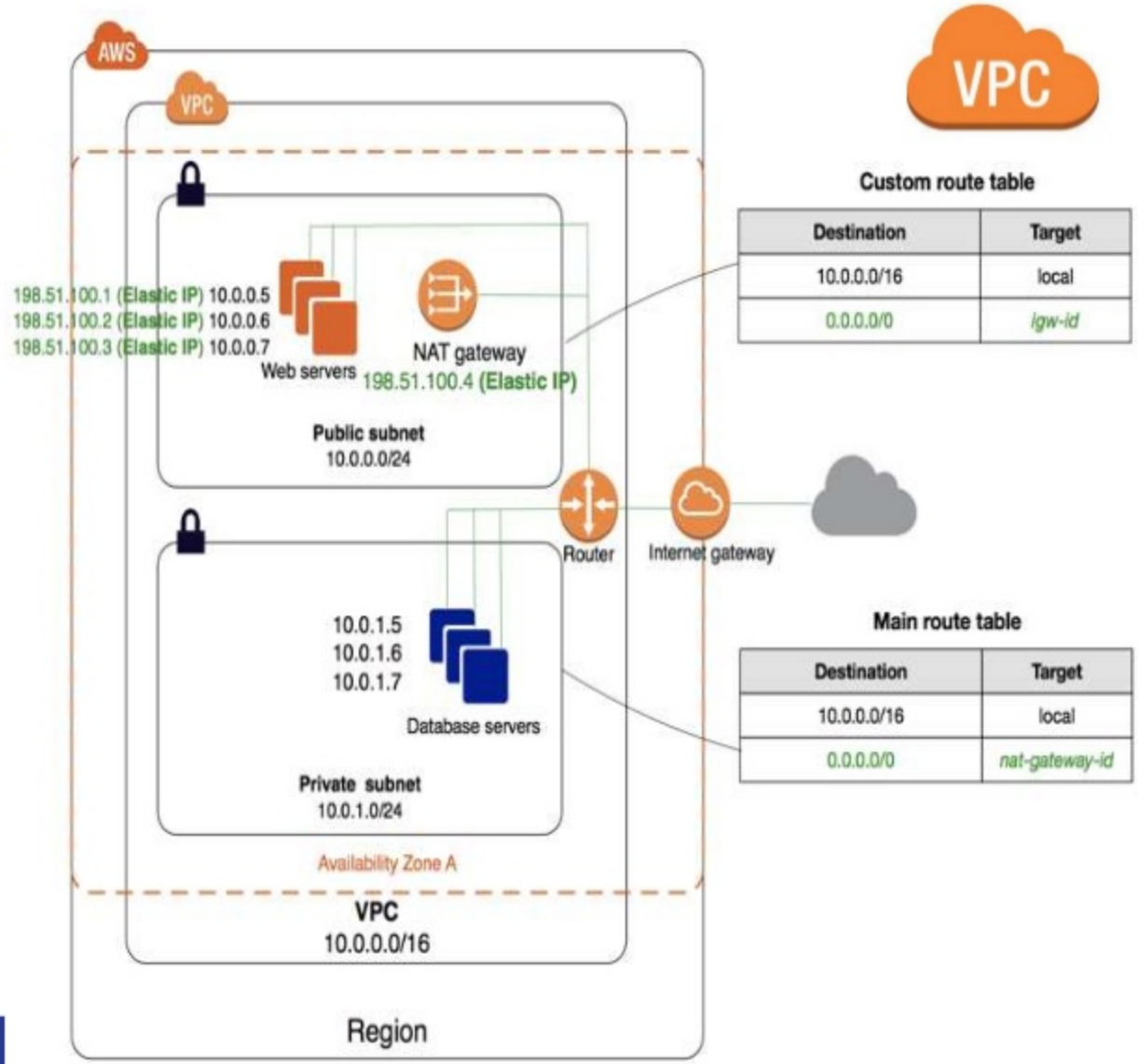


VPN Example


AWS
VPC
VPC
VPC Subnet
VPC Subnet
Availability Zone
Availability Zone
Router
Virtual Private Gateway
VPN
Connection
Customer Gateway
Customer
Network

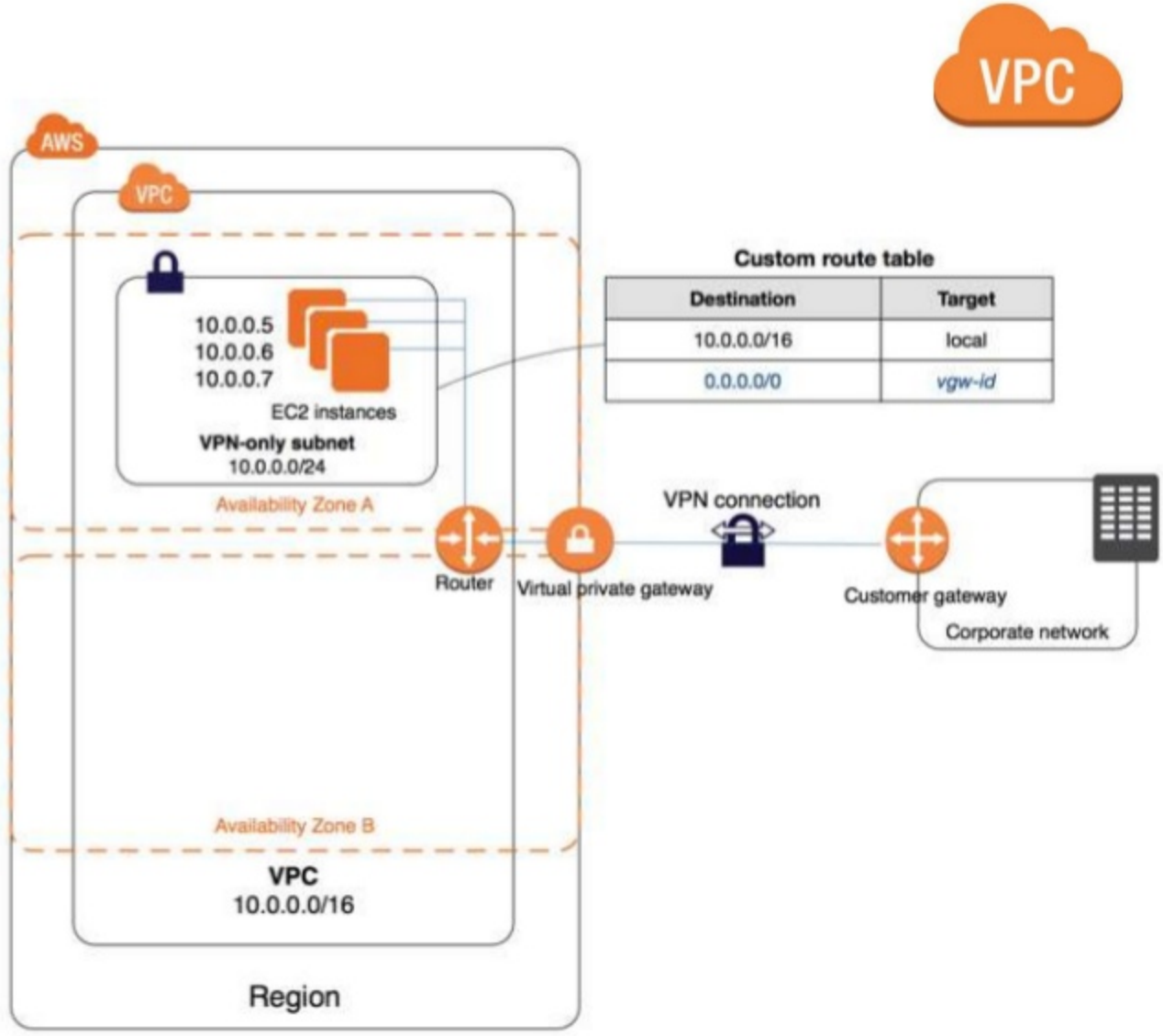
How to Define VPC with private and public subnet

http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/VPC_Scenario2.html



VPC private subnet + Virtual Private Gateway

- https://amazon-aws-big-data-demystified.ninja/2018/06/27/aws-enterprise-grade-networking-security-what-are-your-options-to-protect-your-bigdata/
- http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/VPC_VPN.html
- https://docs.openvpn.net/how-to-tutorialsguides/administration/extending-vpn-connectivity-to-amazon-aws-vpc-using-aws-vpc-vpn-gateway-service/



Data Governance…

Storage Level

Data governance is the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include **availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise** such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be **used by the entire organization.**

S3 coolest options on AWS

- All of your data is replicated on **multi AZ!**
- **Life Cycle** policy
- **Versioning**
- **Cross region** replication
- **Undeletable** bucket - No delete policy
- **MFA on delete Action** - with time interval of 1 min or per action.
- **Deny unencrypted data write**.
- **Analytics**

General Security Concepts | Good to know!

- protecting data while
  - **in-transit** (as it travels to and from Amazon S3) , 2 ways:
    - by using **SSL**
  - **at rest** (while it is stored on disks in Amazon S3 data centers) 2 ways:
    - **Server** Side encryption. (SSE)
    - **client-side** encryption.
  - **In use:**
    - **Hasing….** (dictionary attack?)
    - **Hashing with key**
    - **Any Encryption**

Blog: S3 security options in detail

https://amazon-aws-big-data-demystified.ninja/2018/06/27/aws-s3-security-introduction-and-access-management/

- Detailed Encryption options in AWS
- **Resource based policy VS Identity based policy**

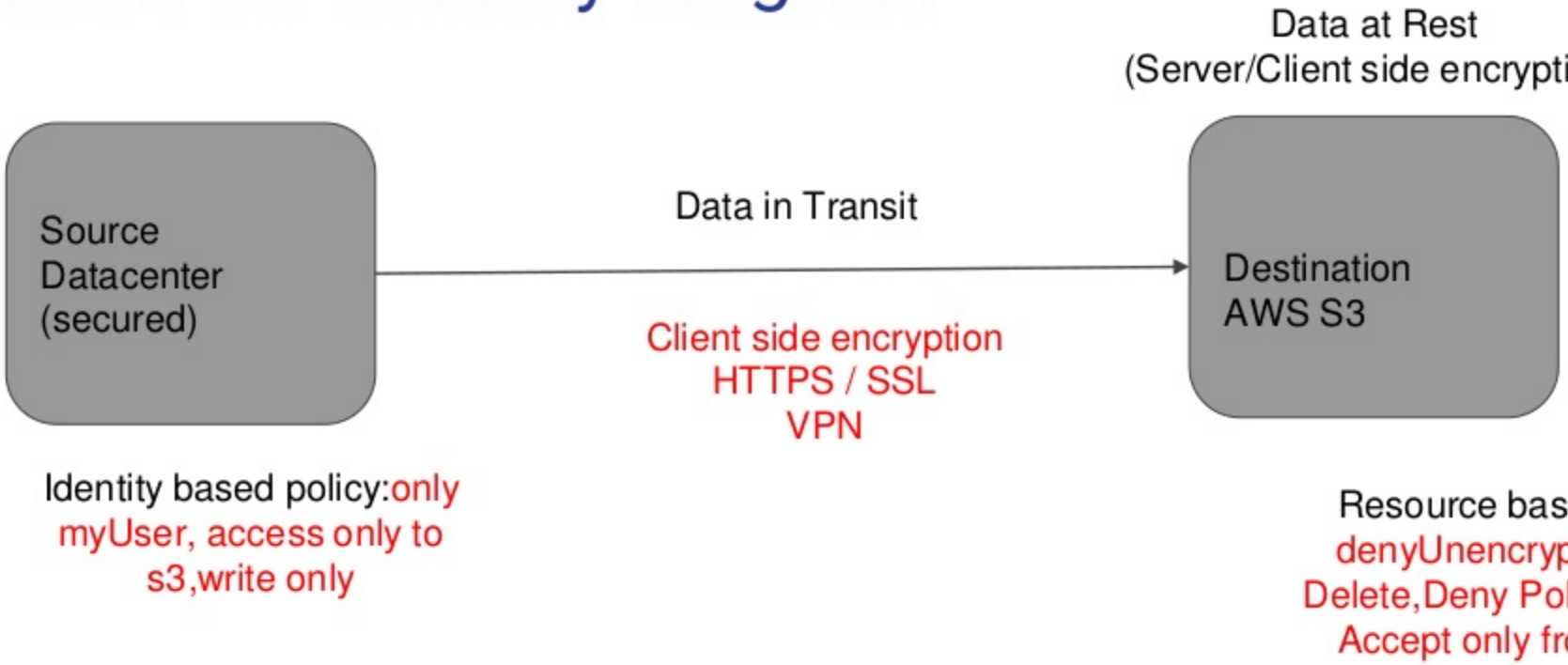
Server Side Encryption (SSE) summary

- **Server-Side Encryption with Customer-Provided Keys (SSE-C)**
  - You manage the encryption keys and Amazon S3 manages the encryption, as it writes to disks, and decryption, when you access your objects
- **S3-Managed Keys (SSE-S3)**
- **AWS KMS-Managed Keys (SSE-KMS)**

Additional aws s3 Safeguard

1. VPN (site to site)
2. IP ACL
3. Identity Based policy (who? Me? S3 read only?)
4. Resourced based policy : e.g deny delete requests/encrypted objects
5. https://amazon-aws-big-data-demystified.ninja/2018/06/27/aws-s3-security-introduction-and-access-management/

Basic S3 Security Diagram



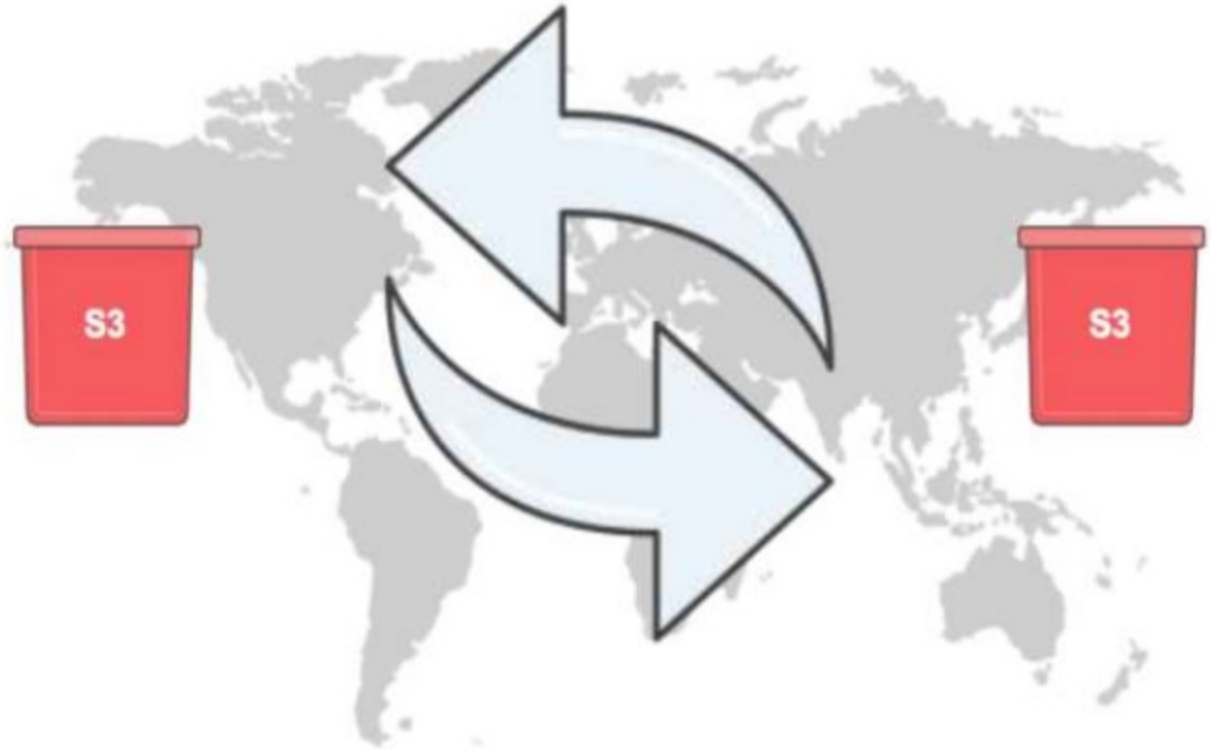
Data Governance…

Data Availability Level

Data governance is the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include **availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise** such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be **used by the entire organization.**

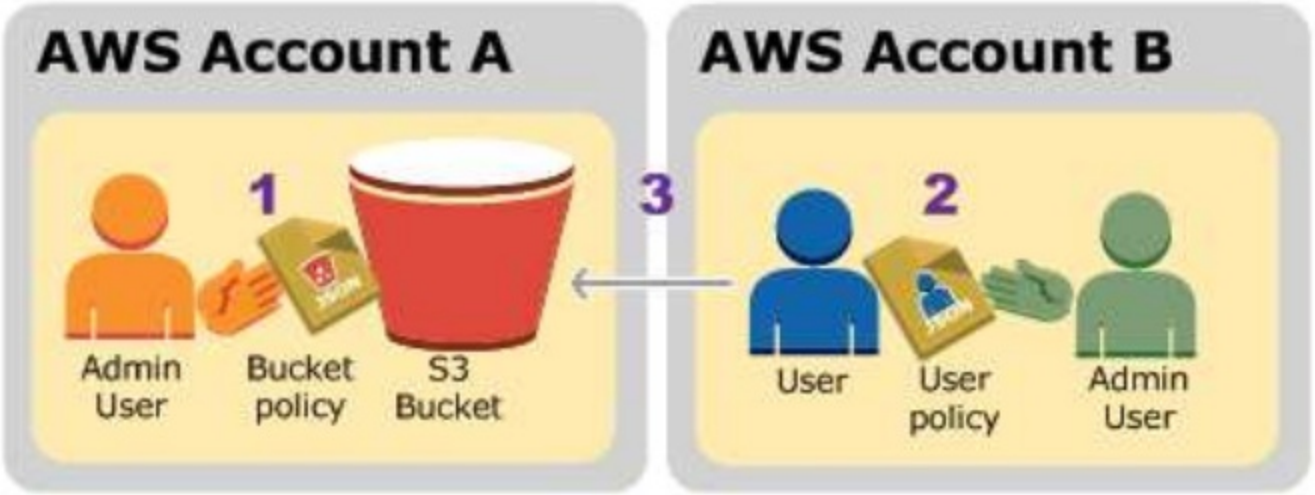
Cross Region Replication

- https://docs.aws.amazon.com/AmazonS3/latest/dev/crr.html
- Versioning must be enabled
- SSL security at transit



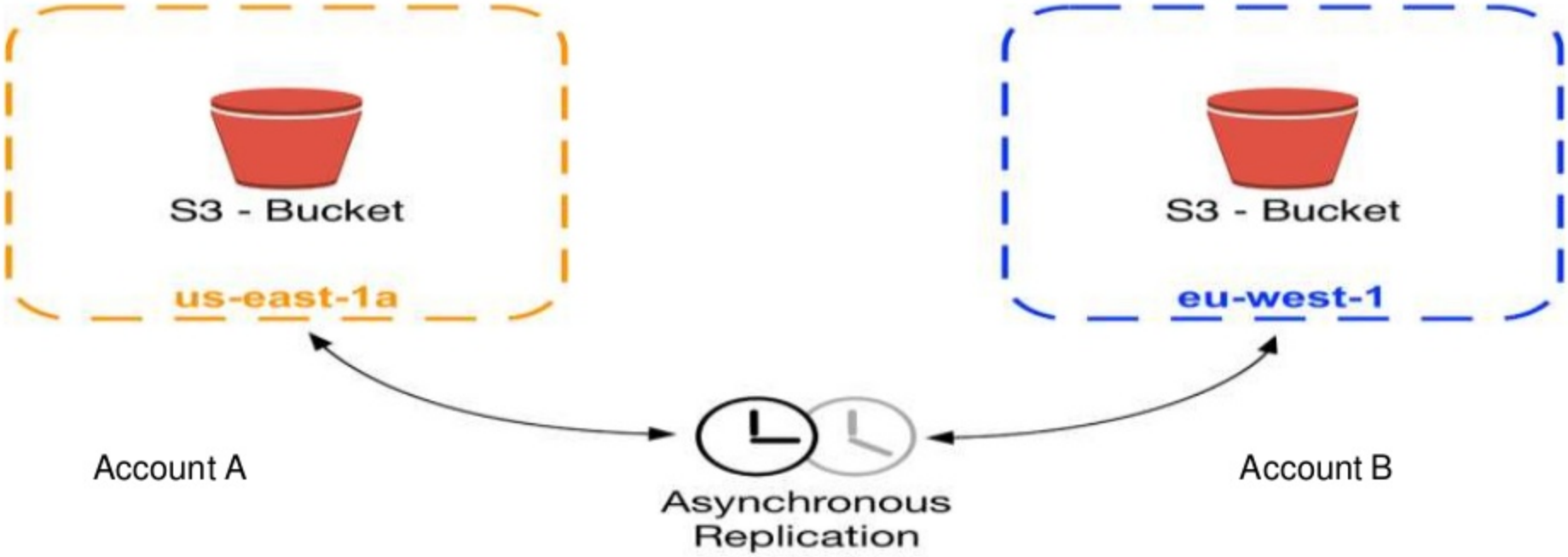
Cross account bucket policy (resource based)

http://docs.aws.amazon.com/AmazonS3/latest/dev/example-walkthroughs-managing-access-example2.html



Cross Region & Cross account replication

https://docs.aws.amazon.com/AmazonS3/latest/dev/crr-walkthrough-2.html



Use Case used by Walla for DR purposes

- Cross account & Cross region replication
  - Copy all backups & and mission critical data
  - Copy all Cloud Formation templates
  - Copy all Code

Data Governance…

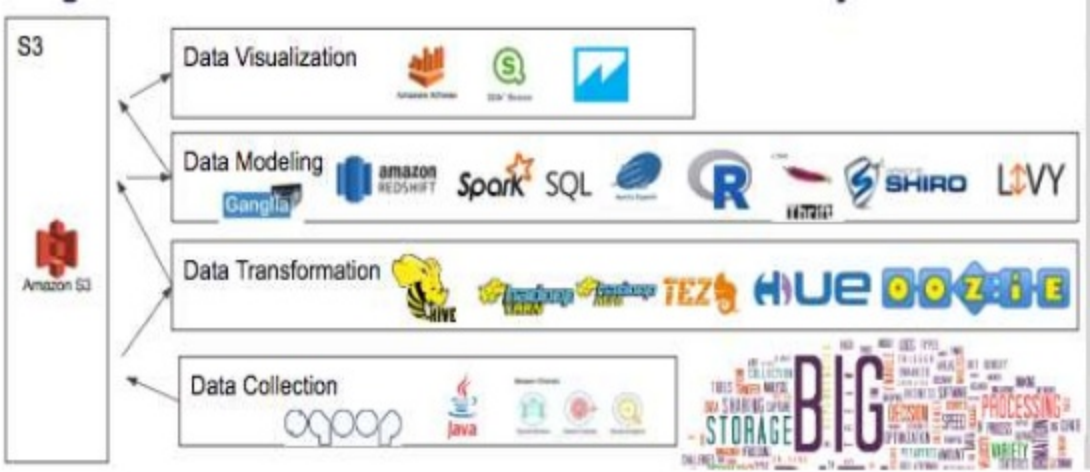
Big Data Security
Architecture Level
(At rest, In motion, In use)

Data governance is the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include **availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise** such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be **used by the entire organization.**

Architecture Security Consideration

- Basicly for each component
  - At rest ,At motion,In use
- Understand different Encryption options per component
  - Prefer SSE
  - Rotate your KMS key X periods
- Restrict access via -
  - Identity based policy: **Role based security** per cluster/technology
  - **Resource** based policy
  - **least privileges**, avoid admin users as much as possible.
  - Manage your EC2 keys wisely
- Understand differences of
  - **Direct Connect VS public internet** VS VPN and their impact on your application
  - **Security groups / NACL / IP Based protection**
  - **private subnet / public subnet on your app**
- For web Consider Cloudfront for GEOlocation protection

EMR Security



- EMR specific:
  - Security configurations
  - Kerberos
- All other webs (Zeppelin, Hue, Oozie)
  - SSL
  - user/password prefer LDAP when possible
  - Shiro use role based access.
- EMR Role → restrict access to s3 buckets
- Identity level - least privileges
- application level (e.g.hive: user level, table level, row level, columns level, kerberos etc)

Data governance…

Privacy in a nutshell: PII
Personal Identifiable Information

Data governance is the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include **availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise** such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be **used by the entire organization.**

PII Privacy challenges in a nutshell

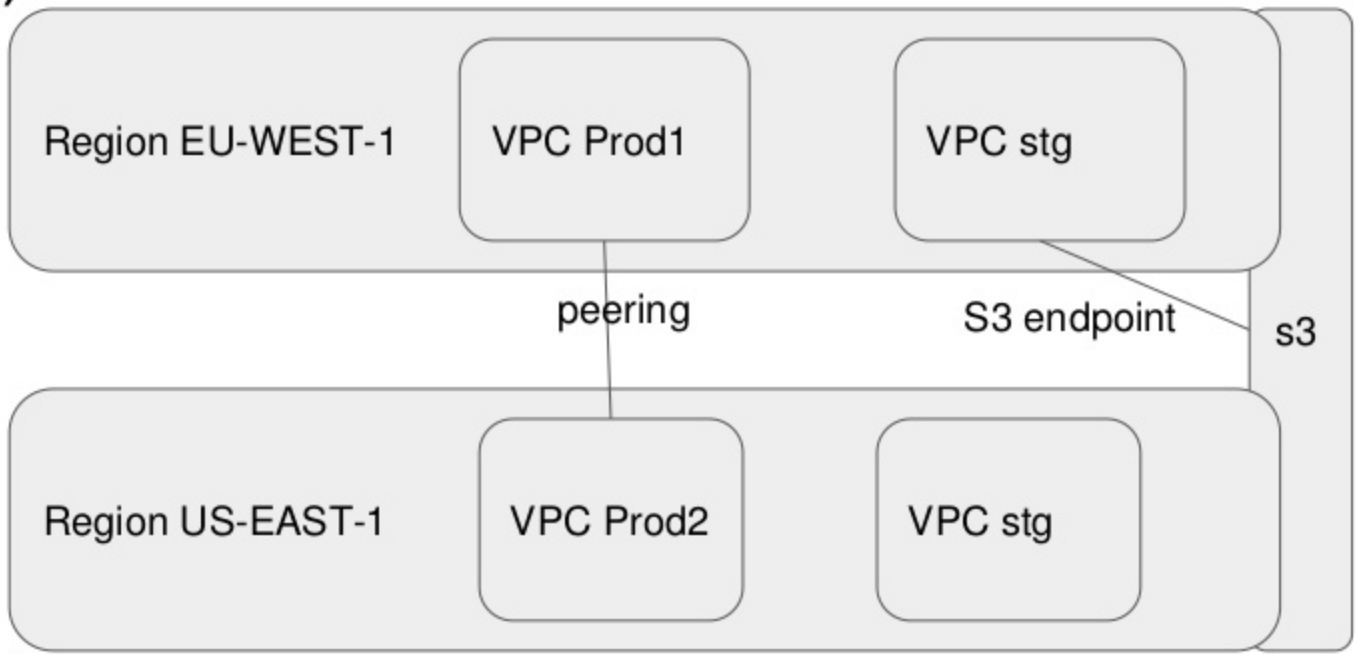
- Problem
  - Assume The attacker has Admin permissions
  - According to Lawyers - no such things a anonymous DB →
    - Your anonymous DB
    - Public 3rd party personal identifiable information DB
    - Joined together… implies your DB is not anonymous.
- **Solution**
  - **Obfuscate & Aggregate** your data wherever possible to avoid a scenario where an anonymous user can be reversed engineered based on the data (location history, habits + timestamp etc)
  - Keep only RAW data you need.

Data governance…

Usability,
& Fine Grained Access Control

Data governance is the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include **availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise** such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be **used by the entire organization.**

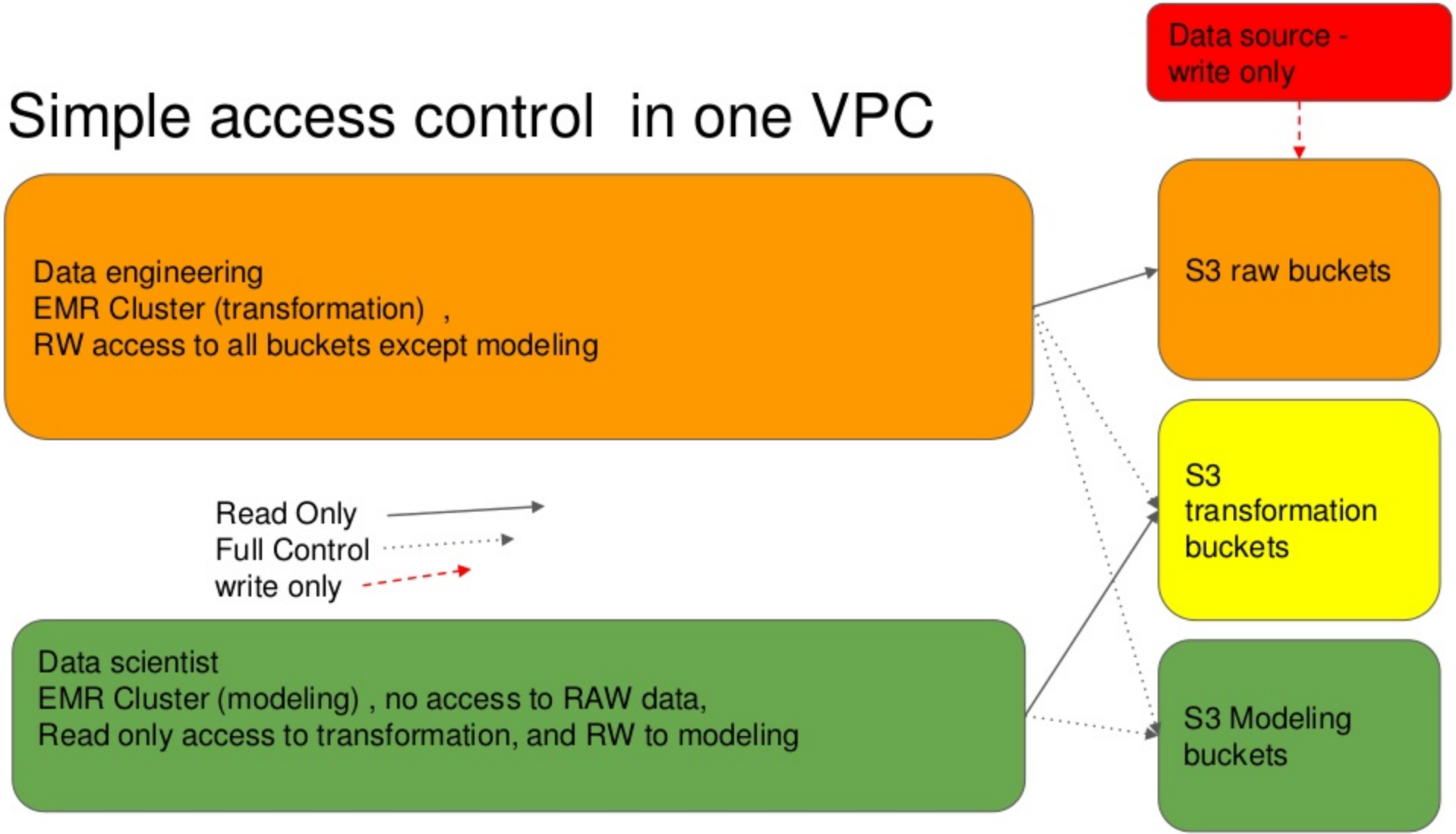
Example AWS Architecture in one account (nothing special)



Account Segregations Motivation

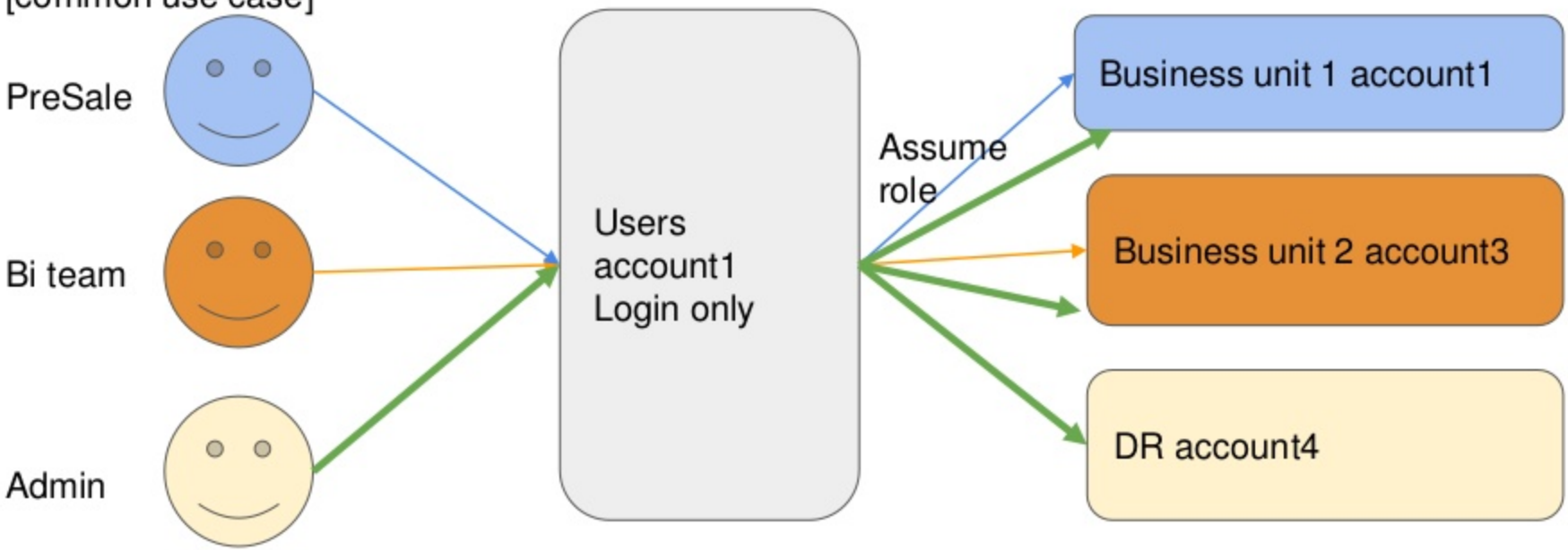
- Fine grained access control.
- Fine grained billing control.
- Limit your Blast radius -
  - what happens if your account is hacked?
  - What happens if one account out of 10 of your accounts is hacked?

Simple access control in one VPC



Simple Account segregation for Business units example

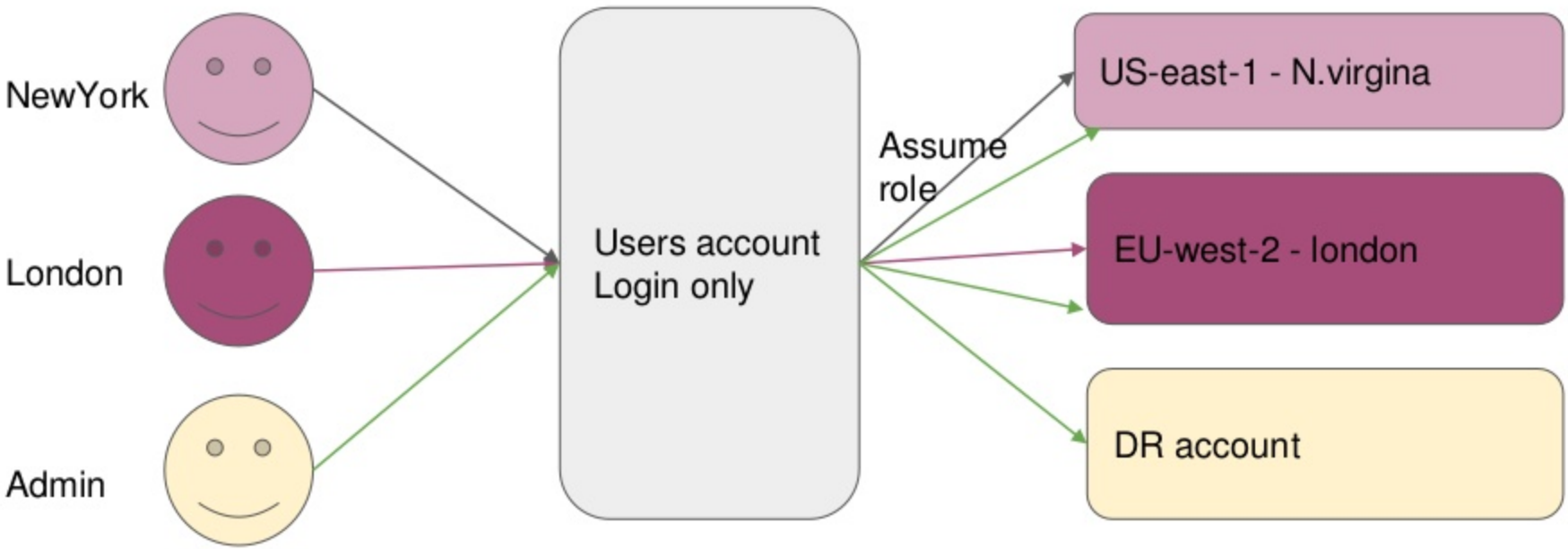
[common use case]



Pre sale and BI may have access to both business units , or just to one business unit,
Admin has access to everything as usual

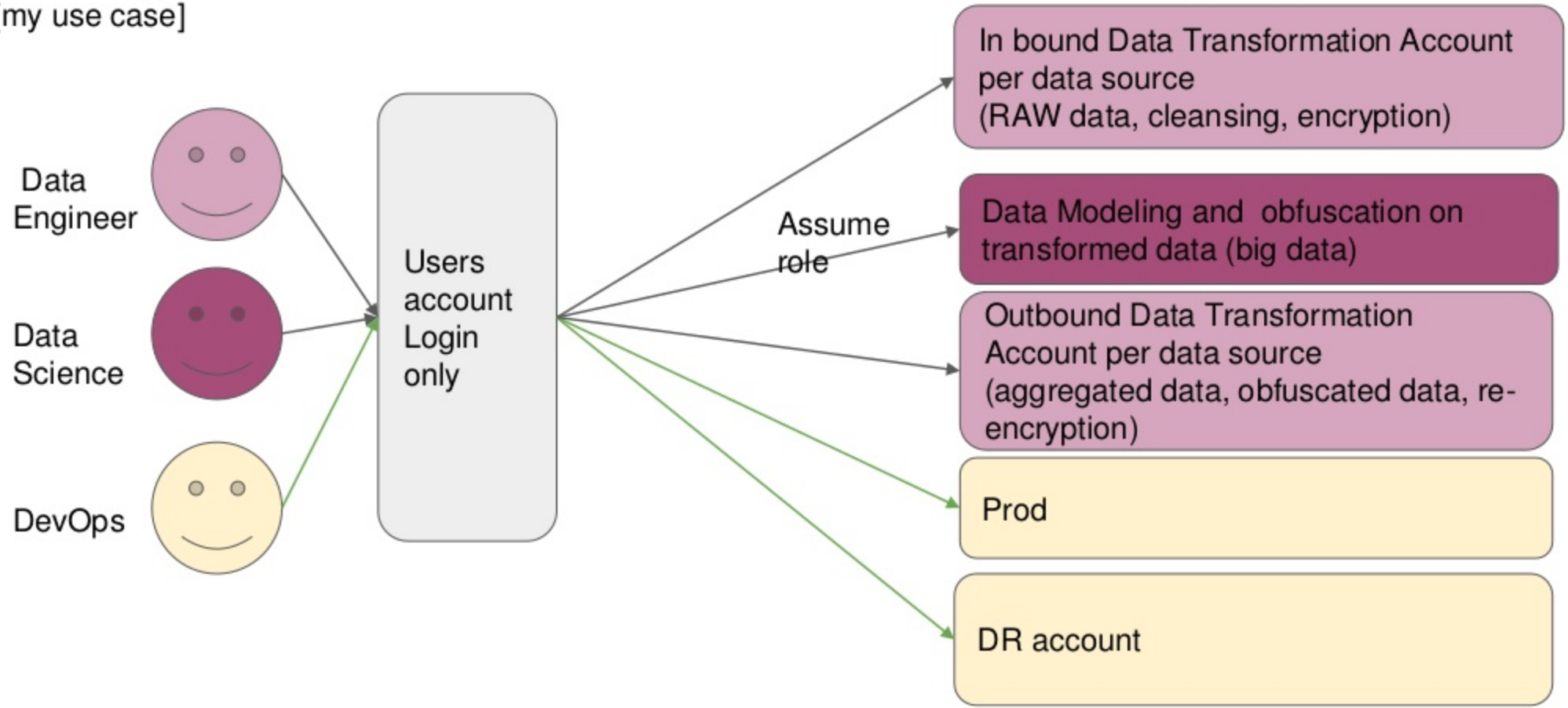
Account segregations per GEOlocation

[data must not leave country use case]

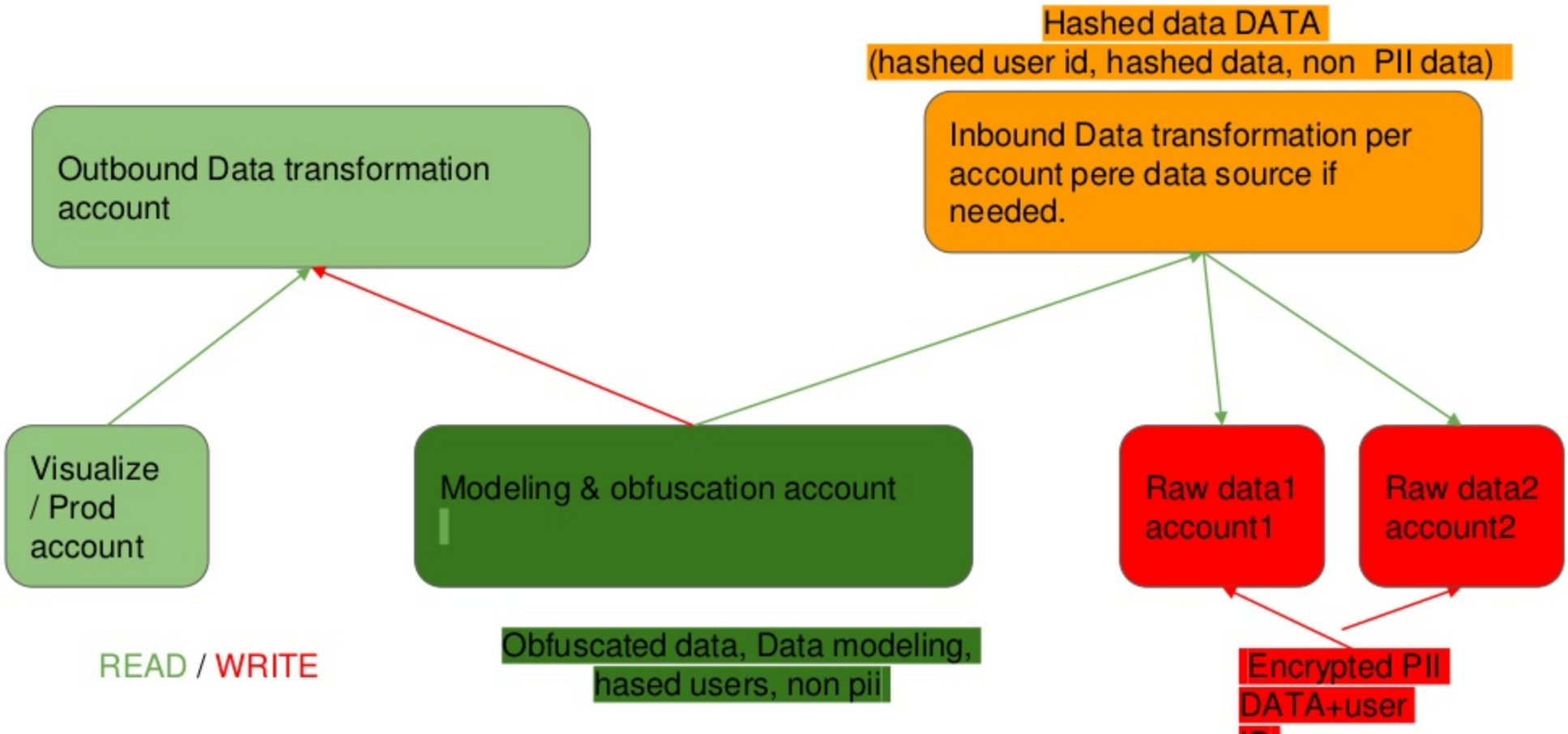


Fine grained access control via Account segregations

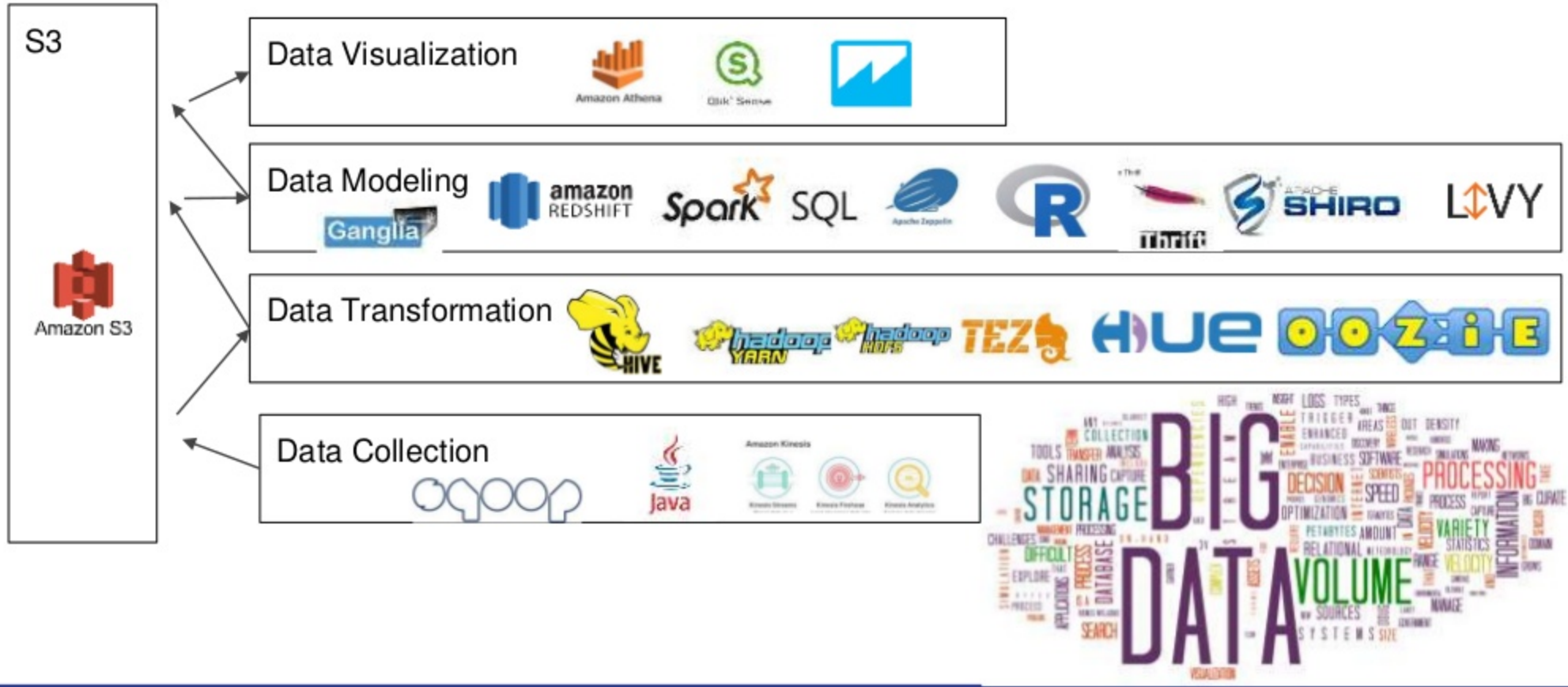
[my use case]



Fine Grained Data Governance Flow via Account segregations



Big Data Generic Architecture | one account



Big Data Generic Architecture | acc. seg.

Amazon S3

Data Visualization

Amazon Athena

Amazon S3

Data Modeling & obfuscation

Spark SQL

Amazon S3

Data Transformation

HIVE

hadoop YARN

hadoop HDFS

TEZ

HUE

OOZIE

Amazon S3

Data ingestion & RAW Data

Amazon Kinesis

Summary and Take messages

- Design you **network from the ground up**, always keep your big data in mind
  - Data access, latency, Data encryption in motion, Performance impact
  - Avoid public IP's when possible. If possible Use Direct Connect / VPN
- Design your **data access layers carefully** per your organization needs.
  - RAW data + encryption @rest & @motion
  - Keep **DR** on data level in mind. [seperate accounts + separate region!]
  - Transformation + hashing level [inbound/outbound]
  - Modeling + obfuscation level
- **Architecture** impact
  - **Security** [at rest, in motion, in use] for each component
  - **Identity** based policy
  - **Resource** based policy
  - Account **segregations**, to avoid blast radius

Stay in touch...

- Omid Vahdaty
- +972-54-2384178
- https://amazon-aws-big-data-demystified.ninja/
- Join our meetup, FB group and youtube channel
  - https://www.meetup.com/AWS-Big-Data-Demystified/
  - https://www.facebook.com/groups/amazon.aws.big.data.demystified/
  - https://www.youtube.com/channel/UCzeGqhZIWU-hIDczWa8GtgQ?view_as=subscriber