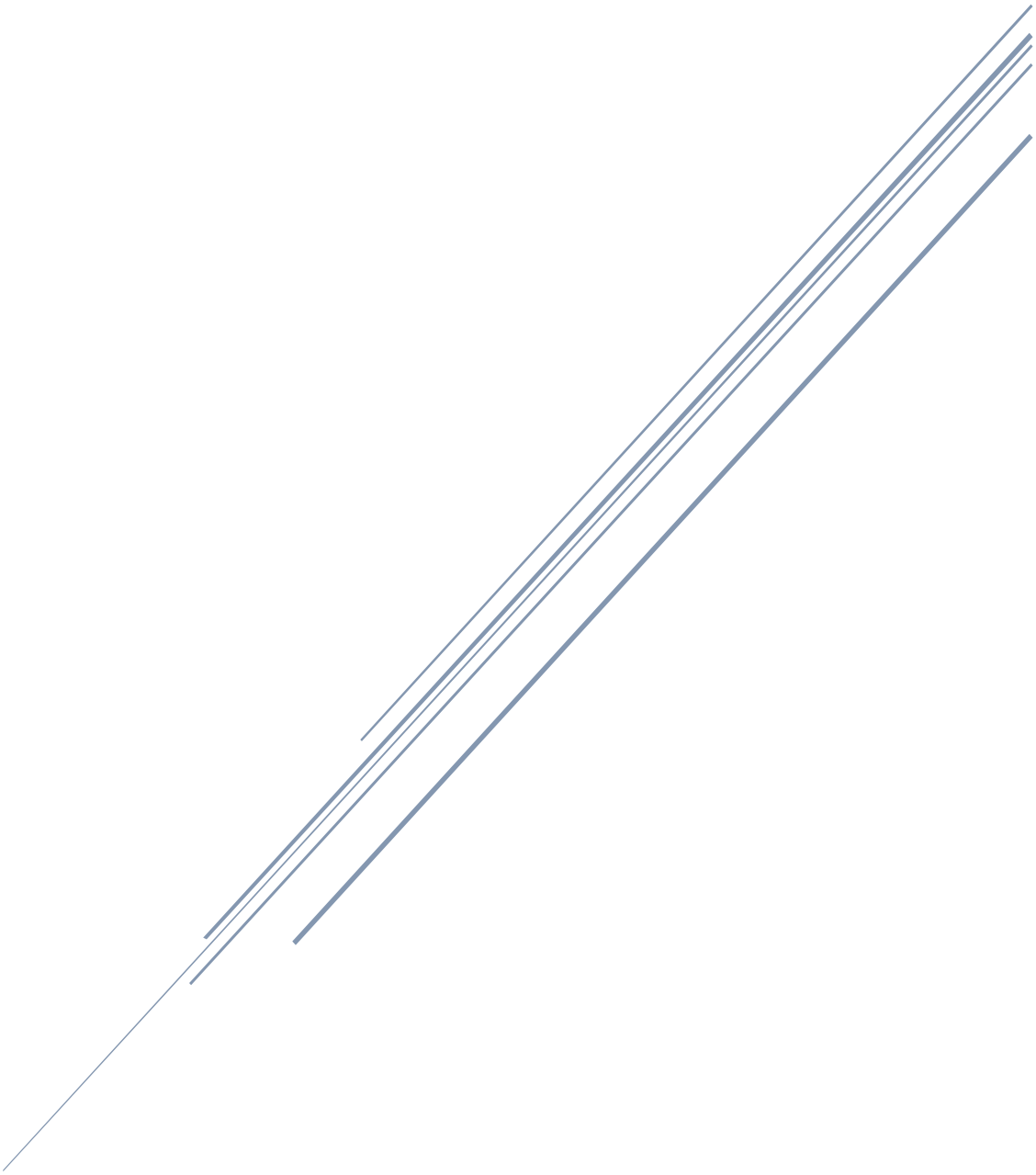
METADATA AWS TARGET ARCHITECTURE DESIGN



Technical Report

Recommended Scalable, Elastic, and Redundant Architecture

Introduction

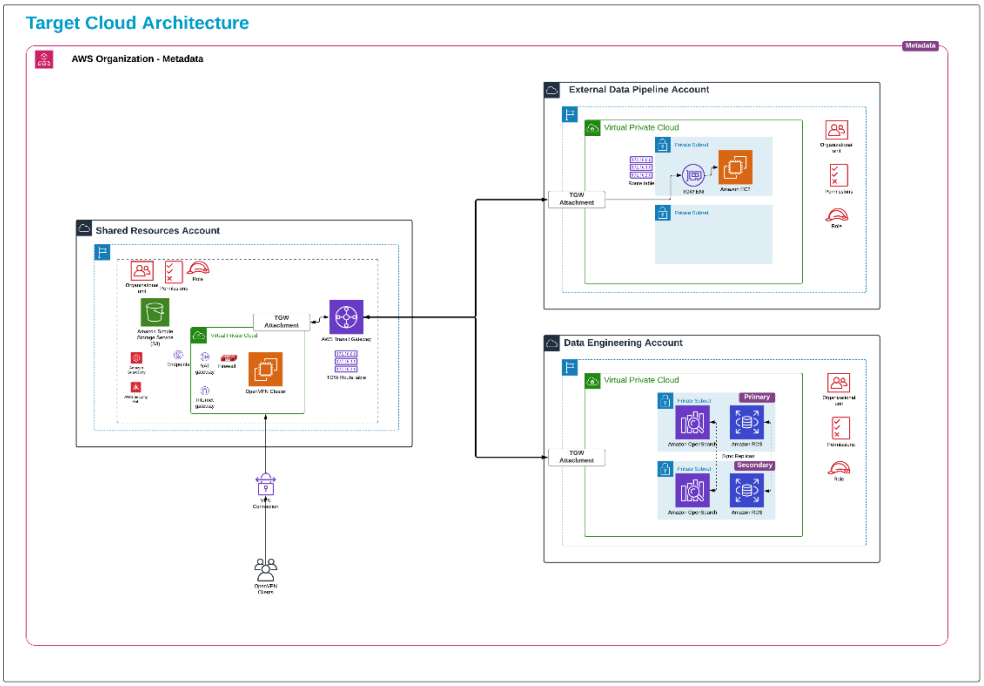
The Metadata team adopted AWS Organizations as a component of a multi-account strategy to achieve twin objectives of business agility and centralized governance as part of their strategic transformation plan. AWS Organizations provides the easiest way to set up and govern a secure, multi-account AWS environment.

AWS Organizations helps companies to centrally manage and govern their environments as they grow and scale their AWS resources. Using AWS Organizations, companies can programmatically create new AWS accounts and allocate resources, group accounts to organize workflows, apply policies to accounts or groups for governance, and simplify billing by using a single payment method for all of the accounts.

The goal is to design a high-level target architecture by organizing workloads in separate accounts and group accounts based on function, compliance requirements, or a common set of controls. At this phase, the plan is to create an AWS Organization with 3 accounts – External Data Pipeline Account, Data Engineering Account and Shared Resources Account. Security and infrastructure scalability were kept in mind to enable common guardrails and service control policies (SCP) be set as the workloads grow. This approach provides boundaries and controls between workloads. Account-level separation is used to isolate these environments, or to provide a strong logical boundary between workloads that process data of different classifications.

Recommended Target Architecture design

The Metadata target architecture utilizes the following account structure to achieve effective security operations and seamless migration of workloads and applications to AWS. These dedicated accounts help ensure separation of duties, support different governance and access policies for different sensitives of applications and data, and help mitigate the impact of a security event.



The above diagram captures the high-level structure of Metadata target AWS account architecture.

All the accounts that are shown in this target architecture are part of a single AWS Organization. The architecture contains the shared resources account, external data pipeline account and data engineering account that house Metadata applications.

At the top level, there are AWS services and features that are designed to apply governance and control capabilities or guardrails across the multiple accounts in Metadata AWS organization (including the entire organization or specific OUs). Service control policies (SCPs), an IAM feature for applying guardrails, AWS CloudTrail for logging all events for all the Metadata AWS accounts, AWS Shield Advanced protection, security groups, and Route 53 are designed as Metadata AWS organization-wide services.

The Metadata Shared Resources account manages the gateway between Metadata infrastructure, applications and the broader internet. It also controls connectivity to the internet and connectivity to the Metadata core network through secure OpenVPN. The Shared Resources account isolates the networking services, configuration, and operation from the individual application workloads, security, and other infrastructure. This arrangement not only limits connectivity, permissions, and data flow, but also supports separation of duties and least privilege for the teams that need to operate in these accounts. This allows network administrators to govern and control networking requirements from a central location.

AWS Transit Gateway (TGW) provides a hub-and-spoke design for connecting all the VPCs as a fully managed service without the provision of virtual appliances. This hybrid connectivity

to a single TWG enables consolidating and controlling the entire routing configuration in one place.

Scenario 1: For EC2 instances from the External Data Pipeline Account to be able to:

a) Read from RDS & OpenSearch instances within Data Engineering Account
b) Write into S3 buckets managed by Shared Resources Account

The architecture diagram depicts how packet flow for the first two steps of the TCP 3-Way handshake between the instances and the bucket.

AWS Security Groups and NACLs are configured to allow communication for the EC2 instances, the RDS and OpenSearch instances.

SYN packet is validated against the External Data Pipeline VPC route table associated to the private subnet of the EC2 instances and the route table has route that points to the TGW attachment. Packet is forwarded to the TGW attachment.

TGW receives the traffic on the TGW route table as the TGW attachment of the External Data Pipeline is associated with the TGW route table.

SYN packet is evaluated against the Data Pipeline TGW route table and forwarded to the Data Engineering account VPC. Then it is evaluated against the VPC route table of the RDS and OpenSearch private subnet of the TGW attachment after the security groups and NACL are evaluated.

For EC2 instances from the External Data Pipeline Account to be able to write into S3 buckets managed by Shared Resources Account, the route table settings to S3 bucket using the Gateway VPC endpoint needs to be configured. S3 bucket policy also needs to be configured to allow access from the gateway VPC endpoint and the Shared Resources Account VPC.

Scenario 2: EC2 instances running within External Data Pipeline Account should allow restricted SSH access.

This can be accomplished either by using secure VPN tunnel and setting up a SCP that only allow restricted SSH access from the OpenVPN cluster.