Cloud Cost Management System for Multi-Cloud Environments

This report presents a comprehensive solution for addressing cloud cost management challenges in multi-cloud environments like Atlan's architecture. The proposed system provides real-time visibility, accurate cost attribution, and clear optimization pathways to prevent unexpected cost increases across AWS, Azure, and GCP deployments.

Major Design Decisions and Trade-offs

Leveraging Existing Infrastructure with Enhanced Capabilities

The solution builds upon Atlan's existing observability stack—particularly VictoriaMetrics, Grafana, and AlertManager—while extending it with dedicated cost management capabilities. Rather than implementing entirely new third-party solutions, this approach reduces implementation complexity and learning curves while maintaining a unified monitoring experience.

Trade-off: While custom development requires more initial effort than using off-the-shelf solutions, it provides greater flexibility and integration with existing systems. The solution might miss some specialized features of dedicated platforms but gains in overall system cohesion and operational simplicity.

Real-Time Cost Data Collection and Analysis

Implementing near real-time cost data collection through multiple channels addresses the delay in cost visibility:

- Direct API integration with cloud provider billing APIs
- Resource utilization metrics as proxies for cost estimation
- Custom collectors for service-specific cost metrics
- Anomaly detection for early warning of unexpected increases

Trade-off: Cloud provider billing APIs typically have some delay (4-8 hours), but this represents a significant improvement over the current situation where delays might span days or weeks. The solution uses resource metrics to estimate costs in the interim, providing timely, if approximate, insights.

Comprehensive Tagging Strategy and Enforcement

A strict, enforced tagging strategy with mandatory tags for business unit, environment, project, and application is implemented to enable precise cost attribution:

- Automated tag validation at resource creation
- Remediation workflows for improperly tagged resources
- Tag-based dashboards and reports

Trade-off: This requires organizational discipline and may introduce friction in resource provisioning. However, the benefits of accurate cost attribution far outweigh these challenges.

Automated Optimization Recommendations

The system implements automated identification of optimization opportunities with clear ROI calculations:

- Idle resource detection
- Right-sizing recommendations
- Reserved instance opportunities
- Spot/preemptible instance candidates

Trade-off: Automated recommendations may sometimes conflict with performance or reliability requirements. The system therefore provides recommendations rather than automatic implementation, allowing teams to evaluate trade-offs.

Proof of Solution

Addressing Delayed Cost Visibility

The solution provides multiple layers of visibility to ensure timely awareness of cost changes:

- Near real-time dashboards showing current resource utilization and estimated costs
- Daily cost summaries with trend analysis
- Anomaly detection alerting for unusual patterns
- Forecasting to predict end-of-month totals

This multi-layered approach ensures that cost increases are detected within hours rather than days or weeks, enabling rapid response.

Solving Cost Attribution Challenges

The comprehensive tagging strategy coupled with detailed resource tracking eliminates guesswork in cost attribution:

- Service-level cost breakdowns
- Team/project attribution views
- Resource-specific cost details
- Historical comparisons

These capabilities allow precise identification of cost increase sources, enabling targeted optimization efforts.

Clarifying Optimization Pathways

The system provides clear guidance on optimization priorities:

- Ranked list of top cost contributors
- Specific optimization recommendations with estimated savings

- Implementation difficulty ratings
- Historical optimization impact tracking

This structured approach eliminates confusion about where to focus optimization efforts, ensuring maximum return on investment.

Known Gaps and Safe Assumptions

Billing Data Latency

Despite improvements, there remains an inherent delay in official billing data from cloud providers. This gap is acceptable because:

- The current situation involves delays of days/weeks, so improvement to hours is significant
- Resource utilization metrics provide reasonable cost proxies in the interim
- The combined approach provides sufficient early warning for most cost issues

Shared Service Attribution

Some costs (like network, shared services) remain difficult to attribute precisely. This limitation is acceptable because:

- The solution uses allocation rules based on usage patterns
- The primary focus is on identifying major cost drivers, which are typically directly attributable
- Even approximate attribution provides valuable direction for optimization efforts

Multi-Cloud Normalization Challenges

Different cloud providers have different billing models, resource types, and optimization opportunities. While the solution normalizes across providers, it may miss some provider-specific optimizations. This trade-off is necessary given Atlan's multi-cloud architecture, and the benefit of a unified view outweighs the loss of some provider-specific insights.