**Enterprise Databricks Capacity Planning & Management Strategy**

**1. Document Control: Initial Draft**

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Date | Author | Description |
| 1.0 | 2025-07-04 | Sri Sivakumar Ramar | Enterprise-scale design & governance |

**2. Executive Summary**

This document articulates the enterprise strategy for capacity planning, performance management, and cost optimization of Databricks as the standard platform for advanced analytics, data engineering, and AI workloads across the organization.

**Objectives:**

* Enable scalable, secure, and cost-efficient processing of large and complex data pipelines.
* Support multi-region operations, compliance with global data residency requirements, and business-critical SLAs.
* Integrate Databricks into the enterprise architecture, governance, and sustainability programs.

**Key Highlights:**

* **Workloads:** Batch ETL (~4 TB/day), streaming (~10,000 events/sec), ML training, and interactive analytics for over 250 concurrent users.
* **Elasticity:** Dynamic autoscaling clusters, resource pooling, and pre-approved quota increases.
* **Resilience:** RTO 4 hours, RPO 1 hour, cross-region disaster recovery.
* **Governance:** Enterprise data catalog integration, role-based access control (RBAC), and regulatory compliance.
* **FinOps:** Proactive cost controls and carbon footprint optimization.

**3. Purpose**

Define a robust, enterprise-grade capacity management framework that:

* Delivers predictable performance at scale.
* Supports high concurrency for diverse business units.
* Provides transparent cost allocation and sustainability tracking.
* Mitigates operational risks across regions and workloads.

**4. Scope**

* Multi-region compute resource planning.
* Storage strategy across hot, warm, and archive tiers
* Integration with enterprise identity, security, and data governance tools
* FinOps and environmental sustainability reporting
* Non-functional requirements for reliability, performance, and compliance

**Out of Scope:**

* Application-specific transformation logic
* Detailed model training pipelines

**5. Background**

This initiative supports the organization’s strategic transformation to a **cloud-native data platform**, consolidating siloed data ecosystems into Databricks as a unified analytics and AI platform. The program encompasses migration of legacy Hadoop workloads, onboarding of new digital services, and enablement of self-service analytics across business units globally.

**6. Enterprise Workload Categories**

|  |  |  |
| --- | --- | --- |
| **Workload Type** | **Description** | **Example Scenarios** |
| Batch ETL | High-volume scheduled ingestion | Daily finance and customer data loads |
| Streaming Analytics | Near real-time processing | IoT device telemetry, fraud detection |
| Interactive Analysis | Ad hoc data exploration | Marketing analytics, product insights |
| Machine Learning | Model training, AutoML experimentation | Personalization, demand forecasting |

**7. Compute Resource Strategy**

**Cluster Modes:**

* **Standard Clusters:** Batch and transformation workloads.
* **High-Concurrency Clusters:** SQL analytics and BI workloads.
* **Single Node Clusters:** Lightweight validation or testing.

**Autoscaling & Elasticity:**

* Min/Max worker thresholds defined per workload class.
* Cluster pools to reduce startup latency.
* Spot VMs leveraged for non-critical workloads.

**Node Types:**

* Memory-optimized (e.g., Standard\_D64s\_v3) for ETL.
* Compute-optimized (e.g., Standard\_F32s\_v2) for ML.

**Headroom:**

* Maintain 20–25% compute headroom to accommodate unplanned spikes.

**8. Storage Architecture**

* **Delta Lake on ADLS Gen2** with tiered retention policies:
  + Raw: 90 days.
  + Curated: 1 year.
  + Archived: 5 years (cold tier).
* Encryption enforced at rest (AES-256) and in transit (TLS 1.2+).
* Cross-region replication of critical datasets.

**9. Non-Functional Requirements (NFRs)**

| **Category** | **Requirement** |
| --- | --- |
| **Performance** | Batch pipelines complete within 2 hours. |
| **Scalability** | Elastic scaling to 2x baseline workloads. |
| **Availability** | 99.9% uptime per region. |
| **Security** | Private VNET deployments with no public IPs. |
| **Compliance** | GDPR, CCPA, and internal data policies enforced. |
| **Sustainability** | 40% of compute on Spot VMs; carbon tracking enabled. |
| **Observability** | Real-time dashboards and anomaly detection. |

**10. Cost Management and FinOps**

**Budgets & Thresholds:**

* Budget alerts configured at 70/90/100% per business unit.
* Quarterly cost reviews with Finance.

**Chargeback Strategy:**

* Tagging clusters (BusinessUnit, Environment, WorkloadType) for allocation.

**Projected Monthly Spend: (samples)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Workload** | **Current Spend** | **6-Month Forecast** | **12-Month Forecast** |
| Batch ETL | $20,000 | $26,000 | $34,000 |
| Streaming | $6,500 | $9,000 | $12,500 |
| Interactive | $4,000 | $5,000 | $6,500 |
| ML Training | $9,500 | $15,000 | $22,500 |
| **TOTAL** | **$40,000** | **$55,000** | **$75,500** |

**Sustainability:**

* Carbon tracking via Azure Sustainability Calculator.
* Cost optimization through Spot instances and auto-shutdown policies.

**11. Governance and Security**

* **Identity Management:** Integration with Azure Active Directory (AAD).
* **Data Catalog:** Unity Catalog for lineage, classification, and RBAC.
* **Network Isolation:** Private Link endpoints, secure VNETs.
* **Audit Logging:** Centralized in Azure Monitor.

**12. Disaster Recovery & Resilience**

| **Failure Scenario** | **Mitigation Approach** |
| --- | --- |
| Regional failure | Cross-region failover and replication |
| Cluster interruption | Autoscaling and retries |
| Data corruption | Delta Lake versioning and snapshots |

**RTO:** 4 hours  
**RPO:** 1 hour

**13. Enterprise Risk Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood** | **Impact** | **Mitigation** |
| Quota exhaustion due to rapid scaling | Medium | High | Pre-approved quota increases, monitoring |
| Cost overruns | High | High | Budget alerts, Spot instances, scheduling |
| Data compliance violations | Low | High | Policy enforcement, regular audits |
| Performance degradation during peak concurrency | Medium | High | Reserved capacity, tuning, scaling policies |
| Regional Azure outage | Low | High | Disaster recovery plan, replication |

**14. RACI Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Activity** | **Data Engineering** | **Enterprise Architect** | **Cloud Admin** | **Finance** | **Security** |
| Cluster sizing & configuration | R | A | C |  |  |
| Storage lifecycle & retention | R | A | C |  |  |
| Budget management & chargeback | C | C | R | A |  |
| Security controls & compliance | C | A | R |  | A |
| Disaster recovery planning | C | A | R |  | C |
| Sustainability tracking | C | A | R | C |  |

**15. Review Checklist**

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Area** | **Checkpoint** | **Status** |
| 1 | Compute Quotas | Pre-approved quotas confirmed |  |
| 2 | Autoscaling | Load tests completed |  |
| 3 | Cost Management | Budget alerts operational |  |
| 4 | Storage | Tiering and retention policies enforced |  |
| 5 | DR | Failover test validated |  |
| 6 | Security | Encryption, VNET, AAD integration complete |  |
| 7 | Governance | Unity Catalog fully operational |  |
| 8 | Observability | Dashboards and alerting in production |  |
| 9 | Sustainability | Spot VMs and shutdown policies active |  |
| 10 | Roadmap | Growth forecast and scaling plan documented |  |

**Final Note**

This Enterprise Databricks Capacity Planning & Management Strategy establishes a comprehensive framework to enable scalable, secure, and cost-effective data analytics and machine learning across the organization.

By combining dynamic autoscaling, proactive governance, robust disaster recovery, and a clear financial accountability model, this design empowers business units to deliver data-driven innovation while maintaining operational excellence.

As workloads grow in scale and complexity, this strategy provides the foundation to:

* Support high concurrency and advanced AI workloads.
* Ensure compliance with global data protection and retention standards.
* Optimize spend through automation, transparency, and sustainability practices.
* Strengthen resilience against operational risks and regional outages.

The principles, controls, and roadmaps outlined here position the organization to unlock the full value of Databricks as a unified analytics platform—driving faster insights, better decisions, and sustainable growth at enterprise scale.