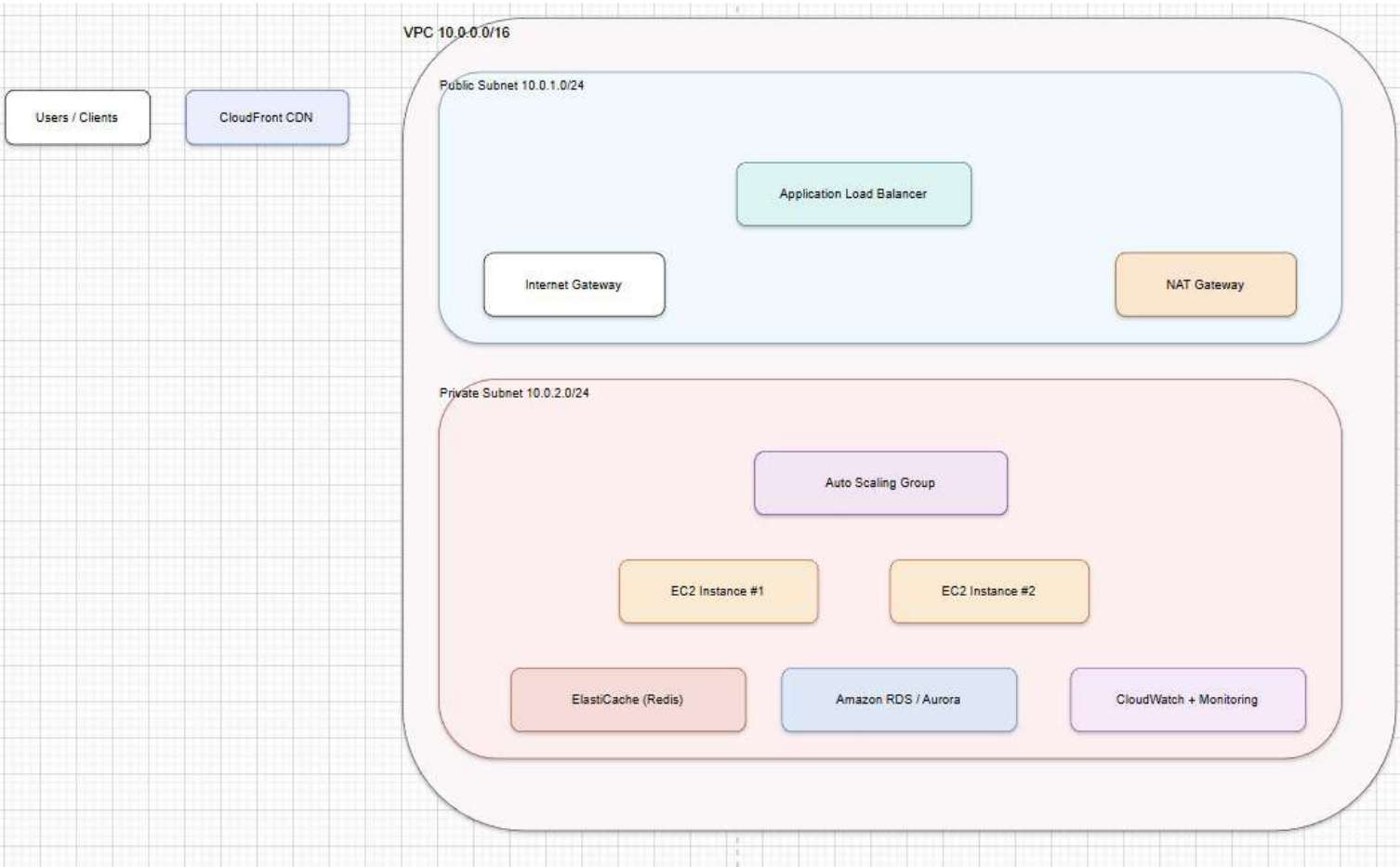
I designed a highly scalable AWS architecture where user traffic first passes through AWS WAF for protection and then reaches an Internet-facing Application Load Balancer deployed in public subnets. The ALB distributes traffic to EC2 instances running inside private subnets through an Auto Scaling Group, ensuring high availability across multiple Availability Zones. Amazon RDS/Aurora is used as the scalable database layer to handle high read and write loads. ElastiCache (Redis) is added to improve performance by caching frequently accessed data. Security is enforced using Security Groups, NACLs, and WAF rules. CloudWatch and Logs provide continuous monitoring, logging, and auto-scaling triggers to maintain performance under 10,000+ concurrent users.


VPC 10.0.0.0/16
Users / Clients
CloudFront CDN
Public Subnet 10.0.1.0/24
Application Load Balancer
Internet Gateway
NAT Gateway
Private Subnet 10.0.2.0/24
Auto Scaling Group
EC2 Instance #1
EC2 Instance #2
ElastiCache (Redis)
Amazon RDS / Aurora
CloudWatch + Monitoring