**InsightPredict Cloud Deployment Options Analysis**

**Overview**

This analysis compares different cloud deployment options for the existing InsightPredict stock recommendation system. The system has already been implemented with a React frontend and requires operationalization for real-world usage and scalability.

**1. Financial API Service Tier Options (Finnhub)**

| **Tier** | **Price** | **Features** | **Limitations** | **Recommendation** |
| --- | --- | --- | --- | --- |
| **Free** | $0/month | - Basic stock data  60 API calls/minute Limited endpoints | - No premium data Rate limiting No support | Suitable only for development and testing |
| **Basic** | $15/month | - 120 API calls/minute Extended data access Basic support | - Limited historical data Some endpoints still restricted | Suitable for MVP with limited users (up to ~50) |
| **Standard** | $29/month | - 600 API calls/minute Full endpoint access Standard support | - Still has rate limitations for high-traffic applications | Good for initial production with up to 500 users |
| **Premium** | $129/month | - 1200 API calls/minute Full historical data Premium indicators Priority support | - Highest tier cost | Necessary for scaling beyond 500 users or requiring premium data |

**Recommendation:** Start with the **Standard** tier ($29/month) for initial production launch. This provides a balance of capabilities and cost, allowing the system to handle approximately 500 users based on the functional requirements. As the user base grows, upgrade to Premium.

**Annual Savings Option:** Finnhub offers annual subscriptions with a discount, which can reduce costs by approximately 10%.

**2. GPU Options for Model Training and Fine-Tuning**

| **GPU Option** | **Configuration** | **Monthly Cost** | **Performance** | **Use Case** |
| --- | --- | --- | --- | --- |
| **AWS g4dn.xlarge** | NVIDIA T4 GPU 4 vCPUs 16 GB RAM | $353/month (on-demand) $161/month (1-year reserved) | Medium performance for training Good for inference | Good for quarterly model updates and daily inference |
| **AWS g4dn.4xlarge** | NVIDIA T4 GPU 16 vCPUs 64 GB RAM | $1,166/month (on-demand) $495/month (1-year reserved) | High performance for medium datasets | Suitable for more frequent model updates or larger datasets |
| **AWS p3.2xlarge** | NVIDIA V100 GPU 8 vCPUs 61 GB RAM | $2,307/month (on-demand) $995/month (1-year reserved) | High-end performance for deep learning | Best for complex models like TFT that require heavy computation |
| **Spot Instances** | Variable configurations | 50-90% discount from on-demand pricing | Same as respective instance types | Cost-effective for non-time-critical batch training |

**Recommendation:** Use a hybrid approach:

1. Use **g4dn.xlarge with Spot Instances** for regular model training, which would cost approximately $65-80/month for 10 days of training time per month.
2. For quarterly full model retraining on the entire S&P 500 dataset, use a **p3.2xlarge** on-demand for a few days (~$230 for 3 days once per quarter).

This approach allows for cost-efficient regular updates while providing the necessary computing power for comprehensive quarterly retraining, meeting requirement F2.5.

**3. Web Application Hosting Options (React Frontend)**

| **Hosting Option** | **Configuration** | **Monthly Cost** | **Features** | **Limitations** |
| --- | --- | --- | --- | --- |
| **AWS Amplify** | Fully managed service | $25-50/month (depends on traffic) | - CI/CD integration Global CDN Easy deployment Built-in security | - Less flexibility for complex configurations |
| **AWS S3 + CloudFront** | Static hosting with CDN | $15-30/month | - Very cost-effective Highly scalable Global distribution | - Requires separate CI/CD setup Manual deployment process |
| **Vercel** | Specialized frontend hosting | $20/month (Pro plan) | - Optimized for React Preview deployments Analytics Easy CI/CD | - Limited control over infrastructure |
| **Netlify** | Frontend-focused platform | $19/month (Pro plan) | - Simple deployment Form handling Identity management Edge functions | - Additional costs for heavy usage |

**Recommendation:** **Vercel** offers the best balance of features specifically optimized for React applications while maintaining reasonable costs. Its CI/CD integration would simplify deployment of the existing React application and provide good scaling capabilities to meet requirements NF1.4 and NF5.3.

**4. Backend API and Processing Infrastructure**

| **Service** | **Configuration** | **Monthly Cost** | **Features** | **Limitations** |
| --- | --- | --- | --- | --- |
| **AWS Lambda + API Gateway** | Serverless compute REST API management | $30-100/month (depends on traffic) | - Pay-per-use model Auto-scaling No server management | - Cold start latency 15-minute execution limit |
| **AWS ECS Fargate** | Container orchestration | $150-300/month | - Predictable performance Fine-grained control No server management | - More complex configuration Higher minimum costs |
| **AWS Elastic Beanstalk** | Managed application platform | $70-150/month | - Easy deployment Managed updates Health monitoring | - Less flexibility Can be more expensive than raw EC2 |
| **AWS EC2 (t3.medium)** | 2 vCPUs, 4 GB RAM | $50-100/month (including load balancer) | - Full control Flexible configuration | - Requires manual scaling More maintenance overhead |

**Recommendation:** Start with **AWS Lambda + API Gateway** for the API layer and data processing components. This serverless approach provides:

1. Cost optimization (only pay for actual usage)
2. Automatic scaling to handle concurrent users (requirement NF1.4)
3. High availability to maintain 99% uptime (requirement NF2.1)
4. Simplified deployment pipeline

As the user base grows, monitor performance and consider migrating compute-intensive components to ECS Fargate if response time requirements (NF1.2) become challenging to meet with Lambda.

**5. Database Options**

| **Database Service** | **Configuration** | **Monthly Cost** | **Features** | **Best For** |
| --- | --- | --- | --- | --- |
| **AWS DynamoDB** | Serverless NoSQL database | $30-100/month | - Auto-scaling Consistent performance Serverless | User preferences, stock metadata |
| **AWS RDS (PostgreSQL)** | db.t3.medium | $150-250/month (with multi-AZ) | - Relational data ACID compliance Strong consistency | Financial data requiring complex queries |
| **AWS ElastiCache** | cache.t3.medium | $100-150/month | - In-memory caching Low latency data access | Caching frequent calculations or API responses |
| **AWS Timestream** | Time series database | $50-150/month | - Optimized for time series Automatic data lifecycle | Historical stock price data |

**Recommendation:** Use a multi-database approach:

1. **AWS DynamoDB** for user profiles and preferences (F4.1-F4.7)
2. **AWS Timestream** for historical stock data (F1.4)
3. **AWS ElastiCache** to improve API response times (NF1.2) and reduce Finnhub API calls

This combination supports the requirements for fast response times while managing costs effectively.

**6. Comprehensive Deployment Tiers**

**Tier 1: Minimum Viable Production (MVP)**

* **Monthly Cost: ~$350-500**
* Finnhub Basic API ($15/month)
* AWS Lambda + API Gateway for backend ($50/month)
* Vercel for React frontend ($20/month)
* DynamoDB for user data ($30/month)
* Spot instances for periodic model training ($50/month)
* AWS S3 for data storage ($15/month)
* CloudWatch monitoring ($20/month)
* **User Capacity: Up to 100 concurrent users**

**Tier 2: Standard Production Environment**

* **Monthly Cost: ~$650-800**
* Finnhub Standard API ($29/month)
* AWS Lambda + API Gateway with higher provisioning ($100/month)
* Vercel for React frontend ($20/month)
* DynamoDB with higher capacity ($60/month)
* AWS Timestream for time series data ($60/month)
* g4dn.xlarge reserved instance for model training ($161/month)
* ElastiCache for response time optimization ($100/month)
* Enhanced monitoring and logging ($50/month)
* **User Capacity: Up to 500 concurrent users**

**Tier 3: Enterprise-Scale Environment**

* **Monthly Cost: ~$1,500-2,000**
* Finnhub Premium API ($129/month)
* ECS Fargate for critical processing components ($300/month)
* AWS Lambda for auxiliary functions ($150/month)
* Vercel for React frontend ($20/month)
* DynamoDB with auto-scaling ($150/month)
* AWS Timestream with extended retention ($150/month)
* g4dn.4xlarge reserved instance for advanced model training ($495/month)
* ElastiCache with higher capacity ($150/month)
* Comprehensive monitoring, logging, and alerting ($100/month)
* **User Capacity: 1000+ concurrent users**

**7. Scaling Strategy and Cost Management**

**Incremental Scaling Approach**

1. **Start with Tier 1** for initial user base (up to 100 users)
2. Monitor the following metrics:
   * API response times (target < 30 seconds per requirement NF1.2)
   * Model accuracy and performance
   * User concurrency and growth patterns
3. **Upgrade components selectively** based on specific bottlenecks rather than moving directly to the next full tier
4. Implement auto-scaling policies for Lambda and DynamoDB to handle traffic spikes

**Cost Optimization Strategies**

1. Use **reserved instances** for predictable workloads (potential savings: 40-60%)
2. Implement **Spot instances** for model training (potential savings: 50-90%)
3. Set up **CloudWatch alarms** for unusual spending patterns
4. Use **S3 lifecycle policies** to archive or delete older data automatically
5. Implement **API caching** to reduce Finnhub API calls
6. Configure **auto-scaling** to scale down during low-traffic periods

**Conclusion and Recommendation**

For the InsightPredict system, we recommend starting with **Tier 1: MVP** and a clear scaling plan to move to Tier 2 as the user base grows. This approach balances:

1. **Cost-efficiency**: Starting with minimal resources while ensuring performance requirements are met
2. **Scalability**: Using serverless and managed services that can scale automatically
3. **Performance**: Ensuring the system meets the 30-second response time requirement (NF1.2)
4. **Reliability**: Achieving the 99% uptime requirement (NF2.1)

The most critical investment areas based on the SRS requirements are:

* GPU resources for model training (to meet requirement F2.5 for quarterly updates)
* Finnhub API tier (to ensure sufficient data quality and API call limits)
* Database performance (to support concurrent users per requirement NF1.4)

By selectively upgrading components based on actual usage patterns, InsightPredict can maintain an optimal balance between cost and performance while scaling to meet growing user demand.