AI, AI, AI... ma quanto mi costi! Come ottimizzare e fare FinOps su soluzioni AI based!

Lorenzo Barbieri @ software one

Microsoft msc TECHNOLOGY NET CODE Azure Meetup TORINO TORINO.NET

AZURE DAY

Platinum Sponsor

Microsoft

msc TECHNOLOGY

Technical Sponsor

.NET CODE

Azure Meetup TORINO

TORINO .NET

stickermule

Packt

JET BRAINS

What is FinOps?

FinOps is shorthand for “Cloud Financial Management” and “Cloud Financial Operations”

It is the practice of bringing **financial accountability** to the variable spend model of cloud, **enabling distributed teams** to adequately position
business priorities between speed, cost, and quality.

At its core, FinOps is a **cultural practice**. It’s the way for teams to optimize and manage their cloud costs, where everyone takes ownership of their cloud usage supported by a central best-practices group.

Cross-functional teams across IT, Finance and Procurement, work together to enable faster product delivery, while at the same time gaining more financial control and predictability.

Why FinOps?

FinOps is about getting the most **business value** from every dollar you spend in the cloud.

On-Premise IT Spend

Decision making team

Procurement | IT & Business Leaders | Finance

- Capital Expense
- Long Planning Cycle
- Formal Purchase Process
- Limited Catalog Items
- Low Pressure for Optimizing

Cloud IT Spend

Cloud Influencers & Spenders

Procurement | IT & Business Leaders | Finance

Engineers

- Operating Expense
- Immediate Infrastructure
- Purchasing at Engineer Level
- Limitless Cloud Choices
- High Pressure to Optimize

Competing KPIs

software one

Cloud Spend and Competing KPIs Iron Triangle

COST

Must meet the Quality and Speed requirements while staying within budget

QUALITY

Release well performing solutions to both internal and external customers

SPEED

Must keep up with the pace of the market, competitors, and internal demand

FinOps Framework

FinOps Framework

FinOps is an evolving **cloud financial management discipline** and **cultural practice** that enables organizations to get **maximum business value** by helping engineering, finance & business teams to collaborate on data-driven spending decision

Principles

- Teams need to collaborate
- Everyone takes ownership for their cloud usage
- A centralized team drives FinOps
- Reports should be accessible and timely
- Decisions are driven by business value of cloud
- Take advantage of the variable cost model of the cloud

Personas

Maturity

Phases

Domains

- Understanding Cloud Usage and Cost
- Performance Tracking & Benchmarking
- Real-Time Decision Making
- Cloud Rate Optimization
- Cloud Usage Optimization
- Organizational Alignment

http://www.FinOps.org

At its core, FinOps is a cultural practice.

↓

Empowerment

↓

Accountability

↓

Collaboration

↓

Adoption and Change Management

Let’s look at the Microsoft world…

A brand new Microsoft whitepaper: FinOps with Azure

Adopting these principles with Microsoft solutions

Microsoft supports these FinOps principles with the following solutions:

Microsoft Cost Management – Consider who from the business, engineering and finance teams needs to be included in budget alerts and what automation is appropriate. Support data visibility to allow collaboration and a centralized team.

Azure Policy – Help enforce organizational standards and to assess compliance at scale. Its compliance dashboard provides an aggregated view to evaluate the overall state of the environment, with the ability to drill down to the per-resource, per-policy granularity. It also helps to bring resources to compliance through bulk remediation for existing resources and automatic remediation for new resources.

Microsoft Power BI – Create, share, and consume business insights in the way that serves you and your role most effectively.

Azure Monitor Workbooks – Similar to Orphaned Resources Workbook and FinOps Insights Workbook, Azure Monitor Workbooks provide great insights on optimizing costs and help with house-keeping of Azure resources which can impact costs.

Microsoft Teams – Integrate the people, content, and tools your team needs to work together effectively.

Azure Advisor – Analyzes your resource configuration and usage telemetry and then recommends solutions that can help you improve the cost-effectiveness, performance, reliability, and security of your Azure resources.

Microsoft Azure learning paths – Learn new skills to boost your productivity and enable your organization to accomplish more with Microsoft Certifications.

But today we’ll focus on the AI part…

…every product has it’s own license/costs/availabilities…

- Per user license
- No additional costs
- Available on certain plans only
- Hourly costs
- Consumption based costs

…let’s deep dive on Azure AI technologies costs!

Azure AI Studio

Unified user interface (UI), SDK, CLI

Azure AI Services	Azure AI Search	Azure Machine Learning
Vision	Retrieval Augmented Generation (RAG)	Ground
Speech	Semantic	Evaluate
Language / Translation	Vector	Deploy
Document Intelligence		Monitor
Azure OpenAI		MLOps / LLMOps
Content Safety		Responsible AI Dashboard

Turing	GPT (too many to list)	DALL-E	Meta Llama 2
Florence	Whisper	Embeddings	Hugging Face

Azure Infrastructure

Azure AI Price Calculator

https://azure.microsoft.com/en-us/pricing/calculator/

- Popular
- Compute
- Networking
- Storage
- Web
- Mobile
- Containers
- Databases
- Analytics
- AI + machine learning
- Internet of Things
- Integration
- Identity
- Security
- Developer tools
- DevOps
- Management and gover...
- Media

Azure AI Bot Service
Create bots and connect them across channels
Add to estimate

Azure AI Search
Enterprise scale search for app development
Add to estimate

Microsoft Genomics
Power genome sequencing & research insights
Add to estimate

Azure Machine Learning
Use an enterprise-grade service for the end-to-end machine learning lifecycle
Add to estimate

Machine Learning Studio (classic)
ML Studio is the GUI-based integrated development environment for constructing and operationalizing Machine Learning workflows
Add to estimate

Azure AI services
Add cognitive capabilities to apps with APIs and AI services
Add to estimate

Azure Open Datasets
Cloud platform to host and share curated open datasets to accelerate development of machine learning models
Add to estimate

Azure OpenAI Service
Apply advanced coding and language models to a variety of use cases
Add to estimate

Health Bot
A managed service purpose-built for development of virtual healthcare assistants
Add to estimate

Azure AI Metrics Advisor

Azure AI Document Intelligence

Azure AI Video Indexer

Chat with Sales

software one

Azure AI Sample Architecture… You should take into account all the components!

[Baseline OpenAI end-to-end chat reference architecture - Azure Reference Architectures | Microsoft Learn]

Azure OpenAI Pay as you Go vs Provisioned Throughput Units

[Azure OpenAI Service provisioned throughput - Azure AI services | Microsoft Learn]

[Right-size your PTU deployment and save big (microsoft.com)]

PayGo	PTU
• Variable cost, usage based	• Consistent cost
• Capped throughput	• Scalable throughput
• No assigned latency	• Assigned latency
Good for: ✓ Sandbox / test environment ✓ Lower volume workloads ✓ Intermittent usage scenarios ✓ Use cases resilient to throughput variability or slowness when service is busy	*Recommended for:* ✓ Production workloads ✓ High volume workloads ✓ Throughput heavy workloads ✓ Latency sensitive scenarios

Azure API Management is your friend!

[Mastering PTU Management for LLMs and GenAI: Optimizing Azure OpenAI for Peak Performance and Cost Efficiency]

Calculating Usage-Based Chargebacks
To calculate usage-based chargebacks for Provisioned Throughput Units (PTUs) when sharing an Azure OpenAI instance across multiple business units, it is essential to monitor and log token consumption accurately. Incorporate the "azure-openai-emit-token-metric" policy in Azure API Management to emit token consumption metrics directly into Application Insights.

Understanding Cloud costs is not enough…

Prioritize FinOps Domains & Capabilities

Capabilities an organization must perform in the practice of FinOps to determine their FinOps maturity level (crawl/walk/run)

Source: data.FinOps.org

FinOps Tooling Challenge

- The Cloud and FinOps market is flooded with tooling. Recent research indicates over **300 tools** claim to have cloud cost optimization
- Organizations lacking a FinOps practice, are often investing time and money in homegrown tools, which **negatively impacts** the ability to effectively manage their cloud spend and value
- Existing tooling may not be delivering desired outcomes or driving business value.

Average

3.7

tools used

Source: data.FinOps.org

SoftwareOne FinOps Mission

“ To accelerate the building of Cloud Financial Management practices and creating long term measurable value from cloud investments.

we're software one.

Driven to deliver technology *and* commercial outcomes

- ✓ Over **10,000,000 users** migrated to the cloud
- ✓ Supporting **6.9 million active users** in their cloud environments
- ✓ **#1 Azure Partner** globally (Azure Expert MSP)
- ✓ **AWS Premier Consulting** Partner
- ✓ **Google Cloud** Premier Partner

- ✓ Global **Microsoft Adoption and Change Management Advanced** Specialisation
- ✓ **World leader in FinOps** (Cloud Financial Management)
- ✓ **Leader in Gartner's Magic Quadrant** for SAM Managed Services
- ✓ **Over 30 years of experience** with commercial models for software & cloud

- ✓ One of the largest certified teams in the world, over 200 FinOps Certified Practitioners
- ✓ Member of the FinOps Governing Board
- ✓ Only provider having a FinOps Certified Platform & Certified Service Provider status as well as leader on the SAM Gartner Magic Quadrant

software one

Thank You!

Vote for my session

Lorenzo Barbieri
Principal Consultant @ SoftwareOne
lorenzo.barbieri@softwareone.com

MVP
Microsoft®
Most Valuable
Professional