How To Actually Make Decisions When Architecting a Data Platform

Ed Freeman

endjin

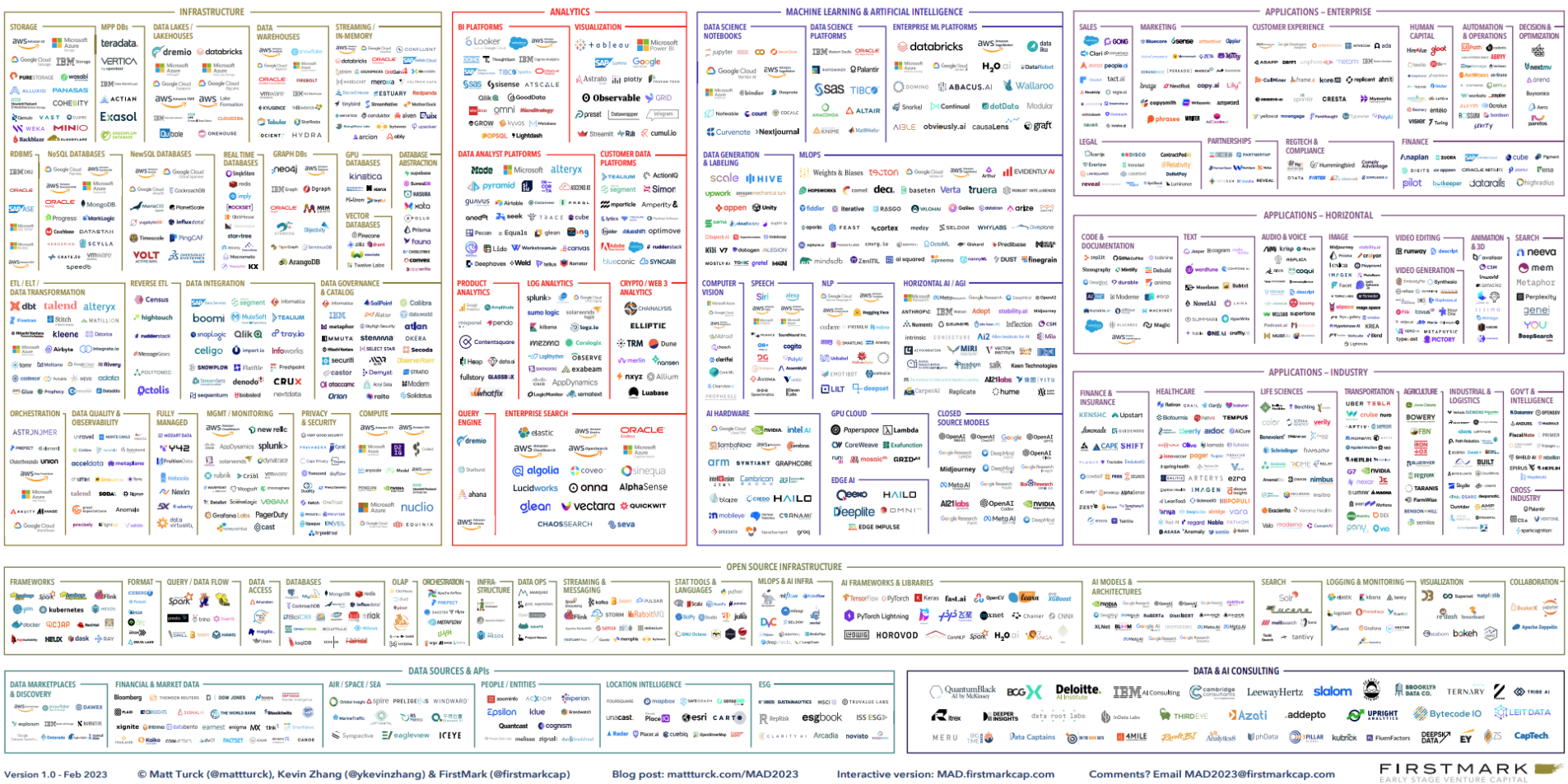
About Me

- Ed Freeman
- Worked at endjin as a Consultant & Data Engineer for 8 years
- Curator of Power BI Weekly
- Dozens of Greenfield and Brownfield projects, across many industries
- Enjoy keeping up to date on latest data trends

endjin

The problem

endjin



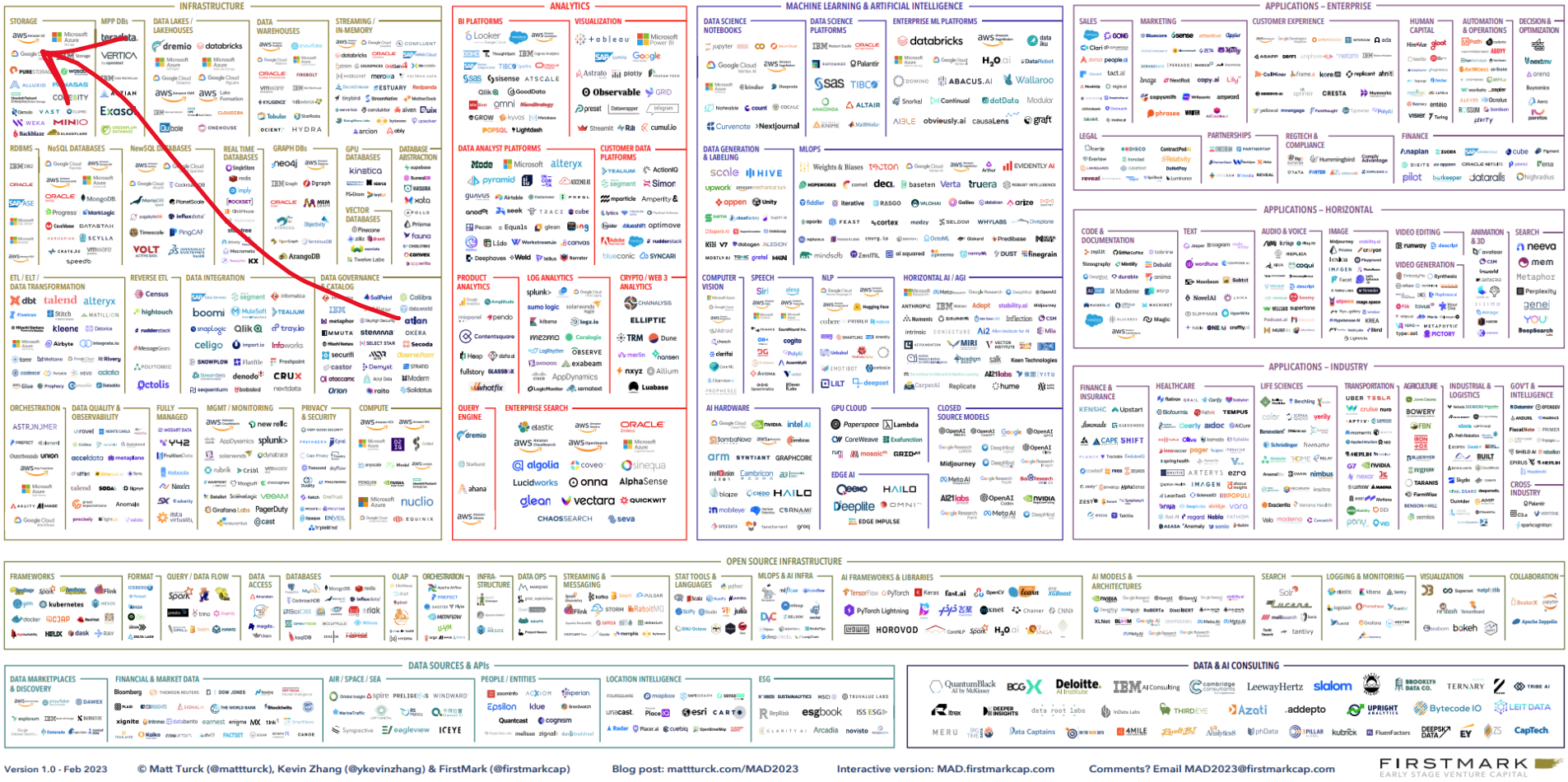
Source: https://mad.firstmark.com/


BIG DATA LDN
21-22 SEPTEMBER 2022
OLYMPIA · LONDON
MISSION TO THE DATAVERSE
databricks
snowflake
Microsoft
Starburst
snowflake
Datoops
Rivery
dremio
databricks
denodo
Collibra
aiven
travel
TigerGraph
THE DATA CLOUD
Datoops Orchestration & Observability
alteryx
coalesce
BIG DATA

Where do I start?

endjin

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



Version 1.0 - Feb 2023

Blog post: mattturck.com/MAD2023

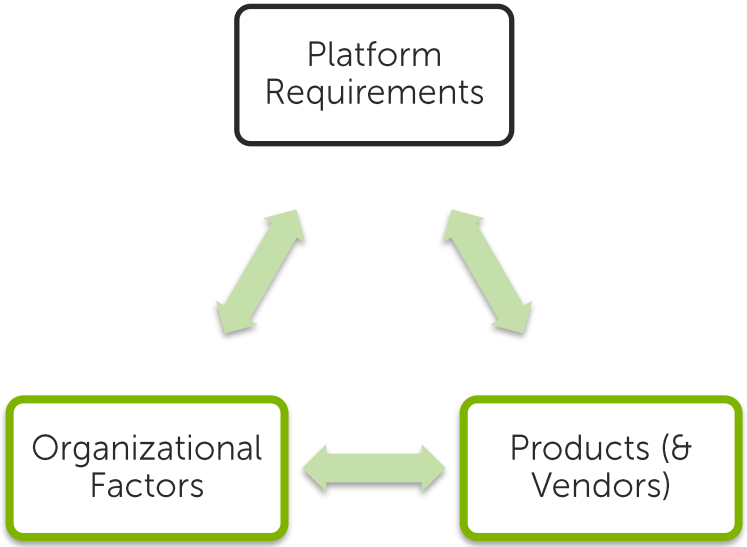
Interactive version: MAD.firstmarkcap.com

Comments? Email MAD2023@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

No but really - where do I start?

- Platform requirements
- Products (inc. Vendors)
- Organizational Factors



endjin

Platform Requirements

endjin

Functional

- What features do I need?
- How should the system behave?
- What are the user flows?

Non-functional

- How should the system perform?
- Where should the system be hosted?
- How should the system be secured?

endjin

Functional – examples

- Platform must support 3rd party databases and APIs
- Data must be stored in an Open Table Format
- System must support SSO
- Business stakeholders can connect to their data models for “self-serve”
- …

endjin

Non-Functional – examples

- System must store 1TB of data
- Platform must cost < $1000 per month
- Platform must be GDPR compliant
- Disaster Recovery RTO of 1 day, RPO of 1 week
- …

endjin

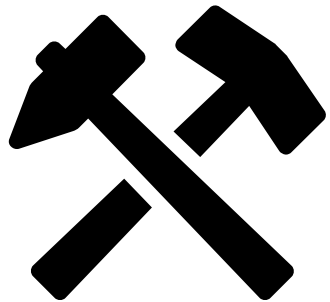
Tools & Methods

Software requirements specification document

- Including Prioritization (use MoSCoW)
- Define user stories
- Ensure requirements are testable
  - BDD (Specflow Online)
- Structured NFR

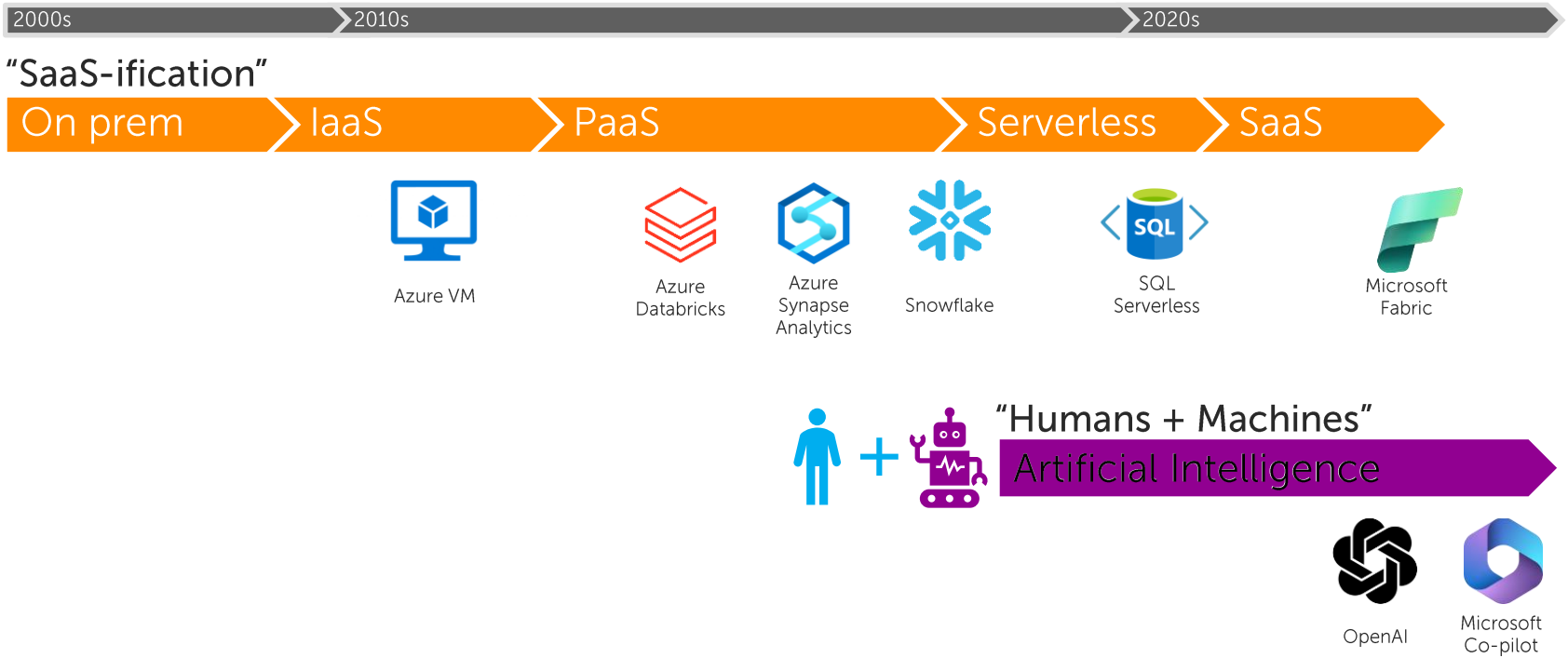
endjin

Products (& Vendors)



endjin

It’s evolving at pace!



Vendors in race to add value. Barrier to entry is lowering.

What core features does it offer?	Is it a unified solution or just one element?	What pricing models does it offer?	What are its integration offerings?	Does it offer free trials?
How mature is the product?	Does it offer substantial learning resources?	How large is the product's community /ecosystem?	What's the product's talent pool for future hires?	What hosting options does it offer?
Does it have AI-enabling features?	Is it a Low/No/Pro code product?	Is the product OSS?	What is their reputation?	What is their funding?
Are they transparent about a product's roadmap?	What is their support model?	Do they embrace/invest in Open Standards /formats?	What is their reputation for LTS products?	Do they have recognised certifications?

Product

Vendor

endjin

General Tips

- Start with subset of products (you already have access to)
- Use highest priority and most constraining requirements
- Attend free training sessions
- Search the internet!
- Take advantage of free-trials
- Perform small PoCs
- Speak to your network
- Review Industry Analyses (e.g. Gartner Magic Quadrants / The Forrester Wave)
- Perform TCO analysis to predict investment

endjin

Product feature sets – promising signs

- Investing in Open Table Formats
- Offering Lakehouse architectures
- Extensive integration with 3rd party data sources
- Supports familiar coding languages
- Have “pro code” integration points (e.g. BYO package)
- AI-enabled functionality

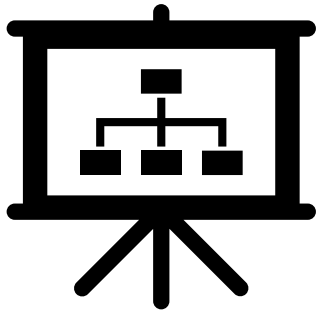
endjin

Careful!

- Best of breed != Best for your overall architecture
- Marketing Hype != Reality
- No roadmap / transparency == Risk
- Industry Analysis eval criteria != your internal eval criteria
- Avoid becoming a “Guinea Pig”
- Avoid the sunk cost fallacy
- Avoid “Tactical” solutions/PoVs becoming production systems
- OSS does not mean “free”

endjin

Organizational Factors



endjin

What other products are being used? Would these be suitable?

Are there any constraining commercial agreements with software providers?

Do you have a CSP aligned with particular products?

What's the organization's risk appetite?

Are data providers in your industry aligning to particular data services?

What are your in-house data engineering skills?

What's your wider organization data culture/maturity?

What's your budget? Capital and Operating.

How does your org handle procurement of new services? How long does this take?

Is there an opportunity to consolidate products?

Does your org have any regulatory requirements?

endjin

General Tips

- Run internal “hackathons”
- Speak to other teams creating data products
  - User groups
- Speak to Procurement / Legal / Finance
- Have a business lead involved throughout
- Think about wider picture within the org
  - Multiple siloed teams using different platforms make it more difficult to enable self-serve
- Think about existing skillsets within your team
  - Are they Data Engineers?
  - “Citizen Developers”?

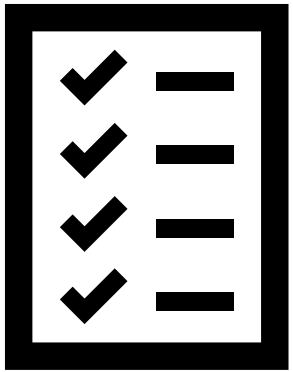
endjin

Careful...

- Internal Politics might make your life difficult
- Your role might be associated with the product you choose

endjin

Decisions, decisions...



endjin

Intoducing... ADRs (Architectural Decision Records)

- Set context for the decision
- List out assumptions & decision drivers
- Highlight and scrutinize considered options
- Capture mitigations for shortcomings of chosen option
- Serve as invaluable history for “future you” and future team members

endjin

Managing ADRs (tips)

- Use versioning
- Treat as append-only
- Establish team of “reviewers”

endjin

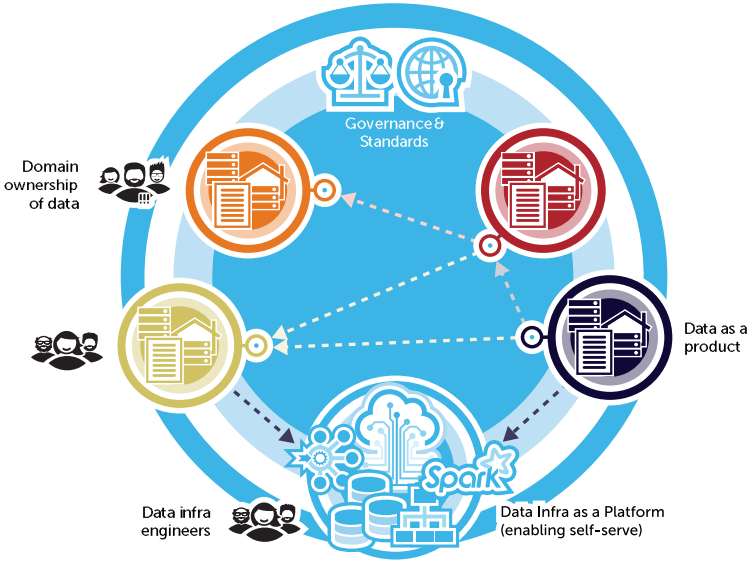
Well Architected Framework (WAF)

- Design Framework to guide the building of high-quality solutions
- Largely platform-agnostic

endjin

Data Mesh

Domain Ownership	• How easy is it to segregate workloads and delegate ownership?
Data Infra as a Platform	• How easy is it to provision new "areas"/"workspaces" for new use-cases?
Data as a product	• How does the platform promote certified/endorsed artifacts and make these discoverable across the org?
Federated computational governance	• How does the platform promote data quality, security, privacy and compliance with regulations?



endjin

Closing Tips

- View endeavour as “socio-technical”
- Upgrade platforms incrementally
  - Take “thin stripe” of functionality to prove value
- Keep architecture as simple as possible (but not simpler!)
  - Keep in mind Conway’s Law & Homomorphic force
- Assess your team’s skillsets (both present and future)
- Determine pricing model best suited with your Operational Budget

endjin

Thank you

ed.freeman@endjin.com (@edfreeman_)

endjin