aws

The Modern Data Community

Contents

This paper provides a perspective on how to build a modern data community for organizations to innovate at scale with data. It covers the challenges in meeting rapidly increasing business demand for data, the bottlenecks in traditional organisational structure and ways of working, and the cross functional talent required in teams.

It describes how to overcome these challenges by decentralizing work, creating autonomous agile teams, building cross functional data skills, and instilling a culture of learning. It addresses how to begin, how to mobilize change, and how to sustain competitive differentiation with data in the long term.

Context

Today, more than ever before, data is the genesis for modern invention. At AWS we see organizations using data to reinvent their businesses, disrupt entire industries, and collaborate across sectors to better serve society. These companies realise that to be successful they must invest in lowering the cost of managing data, unlock business access to data, change mindset in the business, and organize to speed business innovation and value. Forrester research shows that a typical Fortune 1000 will see a $65 million increase in income by increasing data accessibility by 10%.

AWS customers like Intuit have transformed their business with data, Intuit operates 250 data and ML driven products and hold 600 ML patents across their B2B software products for personalised matching of tax experts to clients, affordability predictors for SMB loans, and automated receipt processing. Roche has created a data ecosystem that securely collects sanitized patient and healthcare data from 100 countries for optimized R&D cycles, personalized healthcare, and research collaboration. BMW has partnered with Catena-X, to deliver the first collaborative, open data ecosystem for the automotive industry, linking global players into end-to-end value chains.

Organizations overwhelmingly recognise the importance of data with 99% placing data as a strategic priority. Yet many are struggling to transform with only 48% driving innovation with data, 24% forging a data culture, and 39% managing data as a business asset[i].

As business appetite for data grows, data teams often become a bottleneck, struggling to keep pace with business demand, and slowing the conversion of data into business action. Often, by the time the data team provides an answer to a business question, the question has changed. The continued explosive growth of data adds to the challenge, and data teams cannot maintain a high bar on governing and managing the quality of huge volumes of data, whilst also remaining agile in improving data access to the business. Additionally, business users often retain a legacy mindset, expecting the data team to centrally own all data and serve the business with insight rather than having joint accountabilities. Data skill shortages also persist, making it challenging to scale teams widely across the organisation.

[i] IDC Big Data and AI Executive Survey

Addressing the Challenge

Successful organizations realize that to remove bottlenecks, and empower the business to innovate at scale, requires rethinking how data is distributed throughout the enterprise. These organizations are focused on fostering effective communication and collaboration, and they do this without introducing complex dependencies and hierarchy that slow the movement from data to business action.

> **"** Organizations which design IT systems are constrained to produce **"** designs which are copies of the communication structures of these organisation
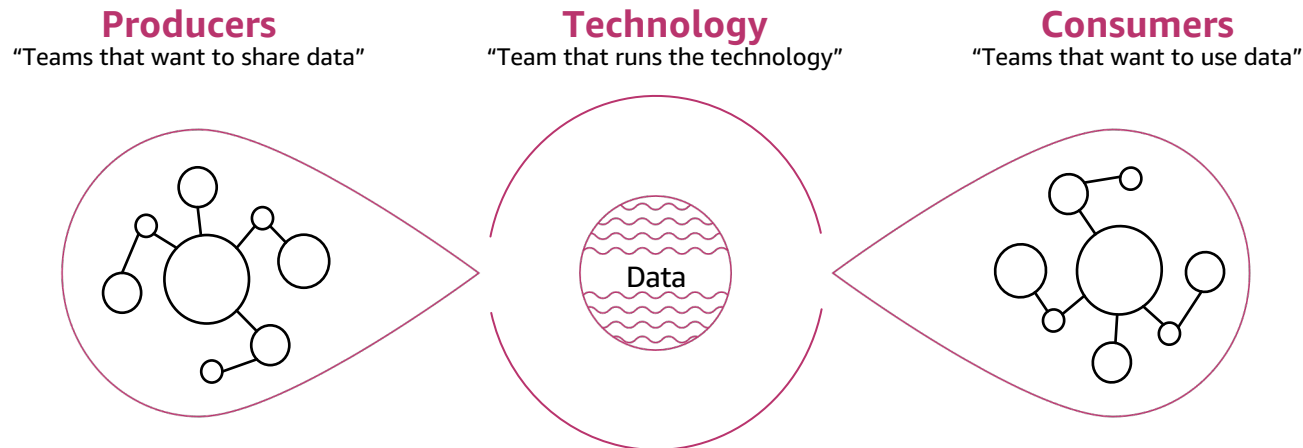>
> Conway's Law, Melvyn Conway (1968)

Conway's Law is a principle that states that the design of a software system reflects the communication structure of the organization that produced it. The law was first formulated by computer scientist Melvin Conway in 1968 and has since become a widely accepted principle in the field of software engineering. Applying this rule to data organisations, the structure of data products will mirror the structure of the organization that built it. If your data organization is a single large department this results in monolithic, slow moving data platforms, if your data organization consists of small distributed teams this results in modular, fast moving data products.
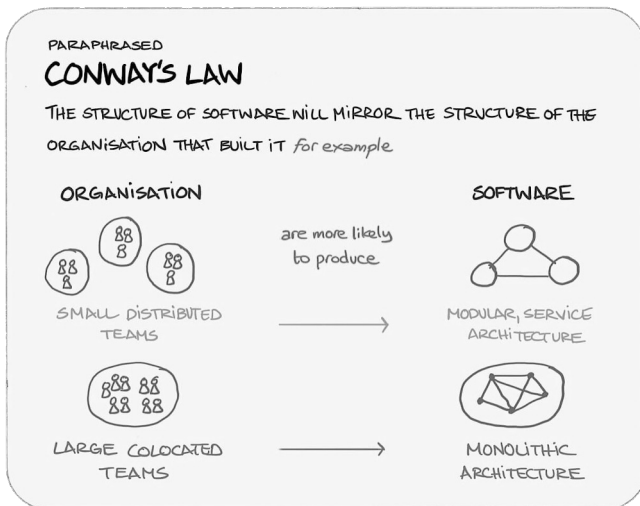
The Modern Data Community

At AWS we refer to this distribution of data work as the Modern Data Community. In the same way that organizations have uncovered benefits by decoupling and moving from monolith IT to micro-services, the modern data community is an organizational and cultural shift from monolithic data organizations to decoupled responsibility. Rather than a single organization (typically IT) being responsible for the ingestion of data, data quality, management of platforms, and the creation of insights, the model pushes responsibility deeper into organizations, thereby increasing autonomy, ownership, and speed.

Producers
"Teams that want to share data"

Technology
"Team that runs the technology"

Consumers
"Teams that want to use data"

Data

Data producers are those teams across the business that want to share data. They are domain experts and are typically aligned to upstream business and application teams. Decisions of data stewardship, and governance largely reside with these teams. They have the deepest understanding of data quality and use metadata tagging, and cataloguing to describe the business and technical attributes of the data. They assign service level objectives (SLOs) to the products they publish.

The **data technology team** is responsible for the operation of the "data ecosystem" by implementing standards and technology to enable the community and its data economy. The platform teams move from traditional operations tasks of monitoring, maintenance, and patching to the value add activities of ensuring the platform meets the business needs and supporting community innovation. Rather than being an innovation bottleneck, this team becomes an innovation enabler.

Data consumers are the teams, individuals, and machines that want to use data for their specific business priorities. Consumers exist across the business in functions such as finance, sales, supply chain, and marketing. They want to minimize non-value add time spent searching for data, requesting access, and testing quality. They want direct access to relevant, reliable, and high-quality data products, and the ability to run analysis in their tool of choice.

Transitioning to a Product Oriented Operating Model

Transitioning to a modern data community requires a mindset shift across the organization to view data as a product rather than as a platform. Companies that treat data like a product can reduce the time it takes to implement it in a new use case by as much as 90%[i].

Amazon.com transformed to become product oriented in the early 2000s as the fast growth of their business saw them go from a small company that sold only books to a multinational, multifaceted organisation. Amazon became aware that as they continued to grow, so too did the complexity, and dependencies of the business, and this was slowing down their pace of innovation.

They realised that they needed to re-invent how they operated to boost agility, remove dependencies, and allow for innovation to be distributed across the organisation. To do this they needed to begin by envisioning the world as a product and break down they way they served customers into modular components or features rather than monolithic platforms. This drove the reorganization of teams away from monolithic departments, and the creation of cross functional, autonomous teams focused on specific products called 2 pizza teams, so named because the teams would be no larger than the number of people that could be feed by two large pizzas.

Restructuring teams short circuited the process of identifying resource and assigning work, rather than bringing the team to the work, work was delivered to the right cross functional team who could manage it holistically with minimal dependencies. Finally work could be deconstructed to take a leaner, agile approach. Executing rapid iterations that reduced risk, and provided feedback faster to inform future direction.
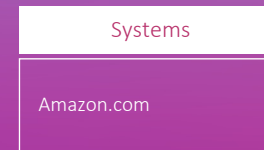
The remainder of this paper details how the principles of a product oriented operating model apply to data, and the importance of these principles in creating a modern data community.
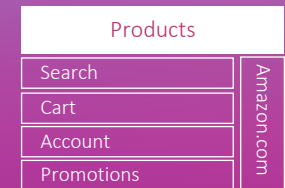
[i]Mckinsey & Co., 2022

FROM   TO
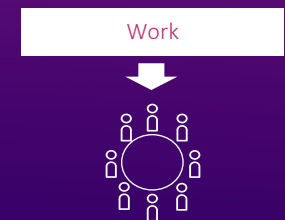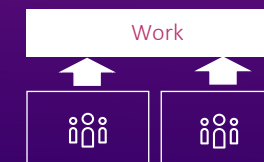
1 Re-envision the world as products

| Systems |
| Amazon.com |

| Products |
| Search |
| Cart |
| Account |
| Promotions |

(Amazon.com)

2 Organize teams around products

| Activity based teams |
| Business | PMO |
| Design | Dev |

| Product based teams |

3 Bring the work to the teams

| Work |

| Work |

4 Reduce risk, move fast through iteration

$

Risk

Jan ──────► Dec

$

Risk

Jan ──────► Dec

Re-envision the world as Products

Traditional, monolithic data departments need to re-envision the solutions they build and the services they provide to structure these as products. Data products need to be shared by both business and IT with the common objective to distribute high quality, trusted data across the enterprise to drive action fast and at scale.

Producers
"Teams that want to share data"

Producer Products

Foundational data products that serve a wide range of business use cases

- Vendor
- Customer
- Employee
- Product

Ecosystem
"Team that runs the technology"

Data Technology Products

Technology products for the Producer and Consumer Communities to leverage

- Data quality and ETL tooling
- Data Catalogue
- Data and ML Ops
- Security controls
- Training

Consumers
"Teams that want to use data"

Consumer Products

Insight, Analytic, ML products to meet business demand

- Customer 360 view
- Financial reporting
- Demand forecasting model
- Ad hoc product analysis

In a modern data community **Data producers** have ownership of foundational data products. Foundational data products serve as building blocks for other data products or analytical solutions within an organization. Data products are high-quality, reliable, and well-documented so that people across an organization can easily access and apply them to a wide range of use cases, from business intelligence to machine learning. Foundational data products often include core data sets, such as customer data, product data, financial data, and operational data. These data sets are used by consumer teams to build analytical solutions that help organizations better understand their customers, improve their products, optimize their operations, and make data-driven decisions.

The **Data Technology** team is focused on building out the technology products that are leveraged by producers and consumers based on the priority demand they see in these communities. This includes products such as a common data catalogue for producers to catalogue their data products, and for consumer to search and subscribe to data products. They build products for data quality management for producers to automate how they clean data and detect sensitive information. The data technology team is also accountable for Data and ML DevOps platforms to standardize the deployment and operations of new data solutions across the business, and they also own training products to help increase data literacy and specialist skills across the organization.

Consumer teams in the business are accountable for creating the insight, analytics, and ML products required to meet with priority challenges or opportunities within their business function. Consumers subscribe to the foundational products created by producers, and they leverage the technology platforms and tools created by the technology team. As consumer teams develop products, these in turn can be published as new producer products.

Organize teams around products:
2 pizza data teams

The 2-pizza team concept emphasizes the importance of small (6-8 people), nimble teams in driving innovation and achieving business goals with data. These are the consumer teams that are embedded into the business, and have a cross-functional makeup including product owners, business analysts, data engineers, and data scientists collaborating to deliver data use cases that meet with business and customer priorities.

2 pizza data teams are autonomous and empowered to focus on priorities for their particular business function. They have limited dependencies on other teams, they break down work and risk to rapidly iterate, and track value as they deliver. Ownership and accountability becomes more natural with each team member assigned a specific role to play and a sense of accountability for the success of the team. These teams are more likely to collaborate and communicate effectively, as team members can more easily stay in sync with each other, and decision making can be streamlined with fewer people involved.

Amazon.com has over 3,300 2 pizza data teams operating across their worldwide business covering their fulfilment centres, Amazon Prime, Amazon Music, Alexa, and Twitch. These teams stay focused on innovating at a rapid clip, using data to deliver compelling customer experiences into their part of the business.

Single threaded leadership

Data Analyst

Scrum Master

Product
Owner

Data Architect

Business
Analyst

Data
Scientist

Data Viz

Data Engineer
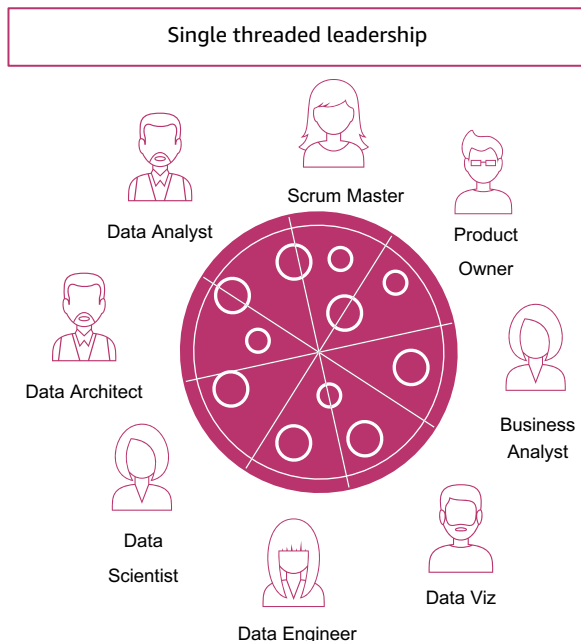
Organize teams around products:
2 Pizza Team Activities and Roles

The roles and activities in a 2 Pizza data team will vary and evolve depending on the maturity of the organization that they are embedded into, and the phase of work across the product lifecycle.

Product Owners represent the Single Threaded Leader of the team. They don't work on anything else and stay focused on defining the vision, prioritizing needs and owning the backlog. The **Scrum Master** works closely with the Product Owner, they operate the scrum and remove roadblocks for the team.

Business Analysts gather requirements, make recommendations for process adjustments, and communicate insights into the business teams. They work closely with **Data Analysts**, who design and maintain data systems, process and clean data in preparation for analysis, and prepare reporting for stakeholders.

Data Engineers develop data pipelines by collecting, transforming, and processing the data needed for business solutions. **Data Scientists** develop, test, and maintain analytic and machine learning models using the data pipelines and feature stores engineered by the data engineer. **Data viz** team members design and deliver the visual and front end aspects of the data solutions. They collaborate with the product owner on user experience requirements, and engage with engineers and data scientists to obtain relevant back end data to feed solutions.

Organize Teams around Products: the modern data community

The concept of 2 Pizza teams can be employed across all components of a modern data community. The roles within each team, the products they develop, and the customer personas they focus on differ, but the attributes of the teams remain consistent, namely to be autonomous, cross functional, and focused on delivering value to their customers to meet with priority demand.

Producers
"Teams that want to share data"
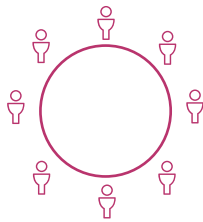
Technology
"Team that runs the technology"
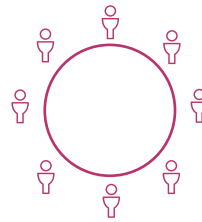
Consumers
"Teams that want to use data"

Primary roles
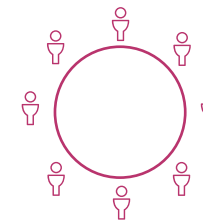
Role	Description	Role	Description	Role	Description
Data steward	• Defines the construct and definitions for data domains	Platform product owner	• Defines the platform vision • Prioritizes producer and consumer needs and owns backlog	Product owner	• Defines the vision • Prioritizes business needs, owns backlog
Data producer	• Manages quality of a data set • Deploys data onto platform • Maintains the schedule and integrity of the data set	Data architects	• Designs data management framework • Defines data architecture standards for production and consumption of data	Scrum master	• Assists on backlog • Operates the scrum, removes roadblocks
				Business analyst	• Gathers requirements • Communicates insights with business teams
Data Engineer	• Collects, transforms, and processes the data needed for data products	Platform engineers	• Develops the solution components for producers and consumers to use the platform	Data Scientists	• Develops, tests, and maintains analytics, AIML products
				Data Engineers	• Collects, transforms, and processes the data needed for business solutions
				Data Visualization	• Designs and delivers the visual/front-end aspects of analytics solutions

Organize teams around products: ML Augmented Productivity

As appetite for data increases in the business the need for specialist data skills continues to grow. A study by the world economic forum shows the top 4 most in demand jobs worldwide are in the fields of Data, Analytics and Machine Learning (ML). This is creating an ongoing skills gap in the market, In the UK 48% of businesses are recruiting for data skills, and 46% have failed to fill roles in the last 2 years[i].

Machine Learning is helping to alleviate the skills gap in data teams. ML enables data teams to work more efficiently and effectively by automating undifferentiated heavy lifting tasks and freeing capacity for tea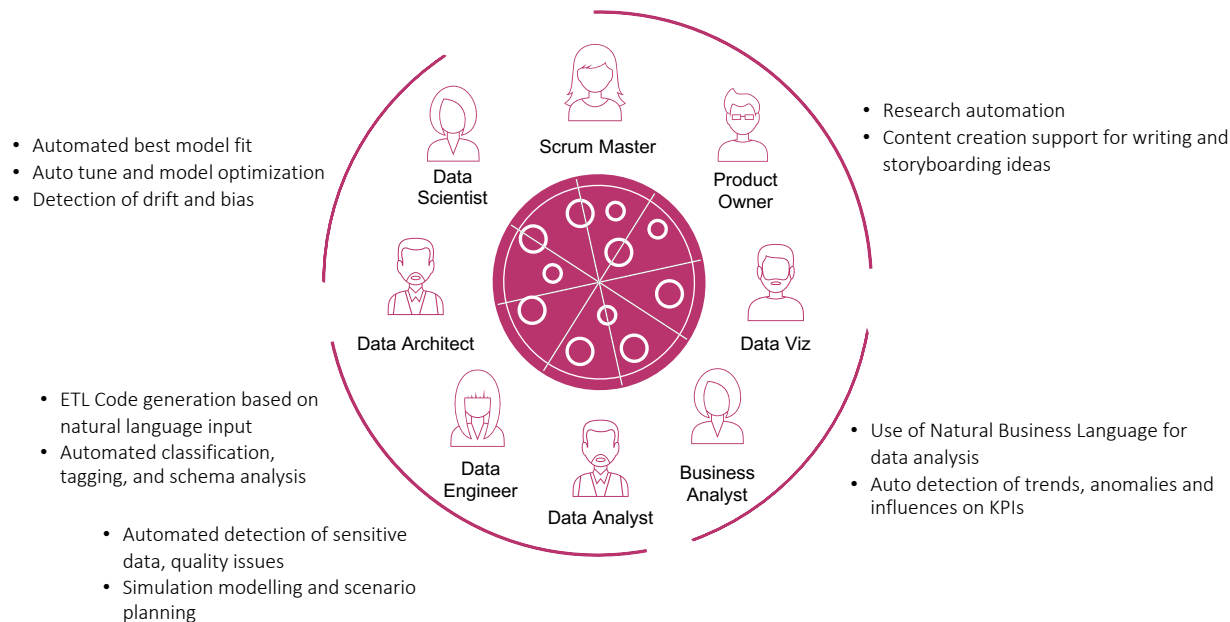ms to focus on high value add activities that continue to deliver differentiation in their business and for their customers. ML also enables wider usage of data across the organization, allowing business users to employ native language tooling to interrogate data, derive insight, and construct narratives without the need for specialist skills. ML is helping data engineers automate the process of data ingestion, cleaning, and transformation. It supports data analysts to identify patterns and trends in data, enabling them to make faster and more informed decisions, and allows business analysts to use natural language tools for data analysis without specialist skills. ML also augments data scientist work, automating the selection of models to fit desired outcomes, tuning models and detecting drift and bias.

With the continued rapid advancement of Machine Learning and AI models the automation bar will continue to rise, eliminating an increasing amount of low value, effort intensive tasks from data teams, freeing up capacity to focus on high value, differentiating work, and enabling business users across the organization to apply data without the need for specialist skills.

- Automated best model fit
- Auto tune and model optimization
- Detection of drift and bias

- Research automation
- Content creation support for writing and storyboarding ideas

- ETL Code generation based on natural language input
- Automated classification, tagging, and schema analysis

- Use of Natural Business Language for data analysis
- Auto detection of trends, anomalies and influences on KPIs

- Automated detection of sensitive data, quality issues
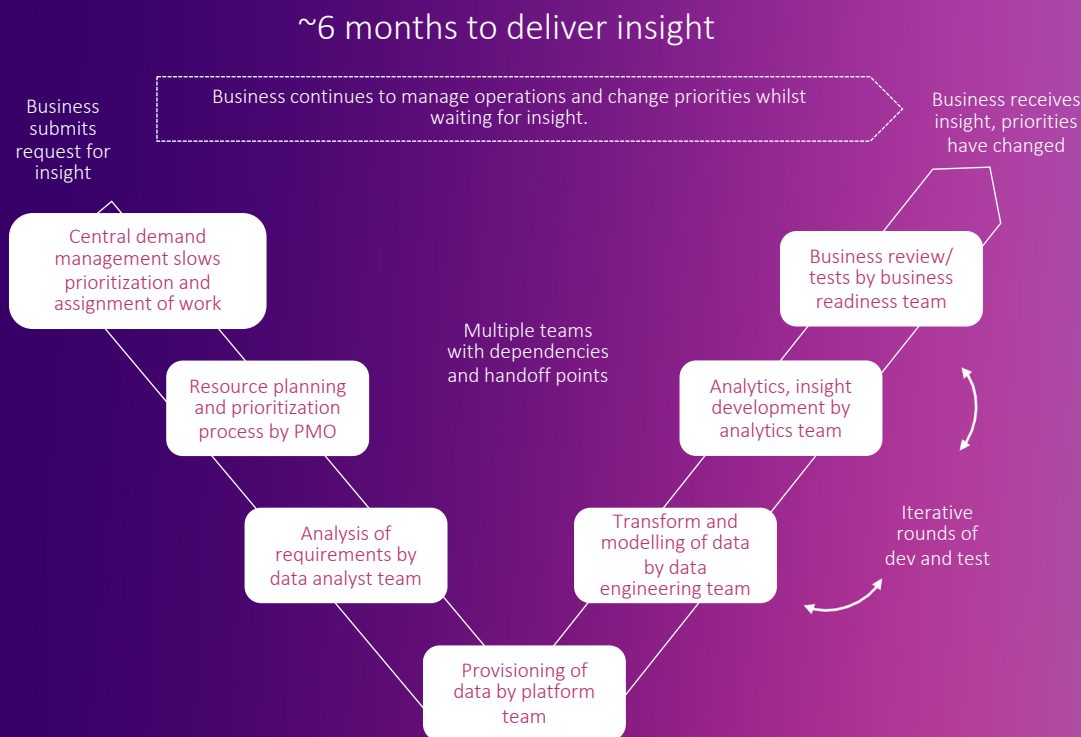- Simulation modelling and scenario planning

[i]Quantifying the UK Skills Gap

Bring the work to the teams: Pain points with traditional delivery

In a traditional organization when the business requires a new data use case this results in a demand request being raised, often through a common demand management process which assesses the request, and assigns the right resource to analyse, prioritise and assign the work. Doing so requires identifying right talent, coordinating availability, estimating effort, and ring fencing time to be able to focus on the work. This process is time consuming and complex with individuals across many different teams needing to be coordinated, scheduled and approved to take up the work even before it can begin.

When the work does begin, the execution requires careful lining up of chained dependencies across multiple teams. This begins with deeper discovery of requirements by the data analysis team, provisioning of the right data into a dev environment by the platform team, transforming and modelling the data by the engineering team, developing analytics with the analytics team, and testing with the test team. This linear development process creates siloed accountabilities with work thrown over the wall. It also limits involvement from the business, raising the risk of wrong understanding of needs, and poor buy in to the end solution. All of this results in slow time to delivery of new insights (typically 6 months) and often by the time insights are delivered, the business need has changed.

Structuring talent into 2 pizza data teams breaks the bottlenecks monolithic departments bring. But placing the right people together in a single team is not enough. These teams need to take a minimal viable product approach, and iterate fast to maximise velocity of value back to the business.

~6 months to deliver insight

Business submits request for insight

Business continues to manage operations and change priorities whilst waiting for insight.

Business receives insight, priorities have changed

Central demand management slows prioritization and assignment of work

Business review/ tests by business readiness team

Multiple teams with dependencies and handoff points

Resource planning and prioritization process by PMO

Analytics, insight development by analytics team

Analysis of requirements by data analyst team

Transform and modelling of data by data engineering team

Iterative rounds of dev and test

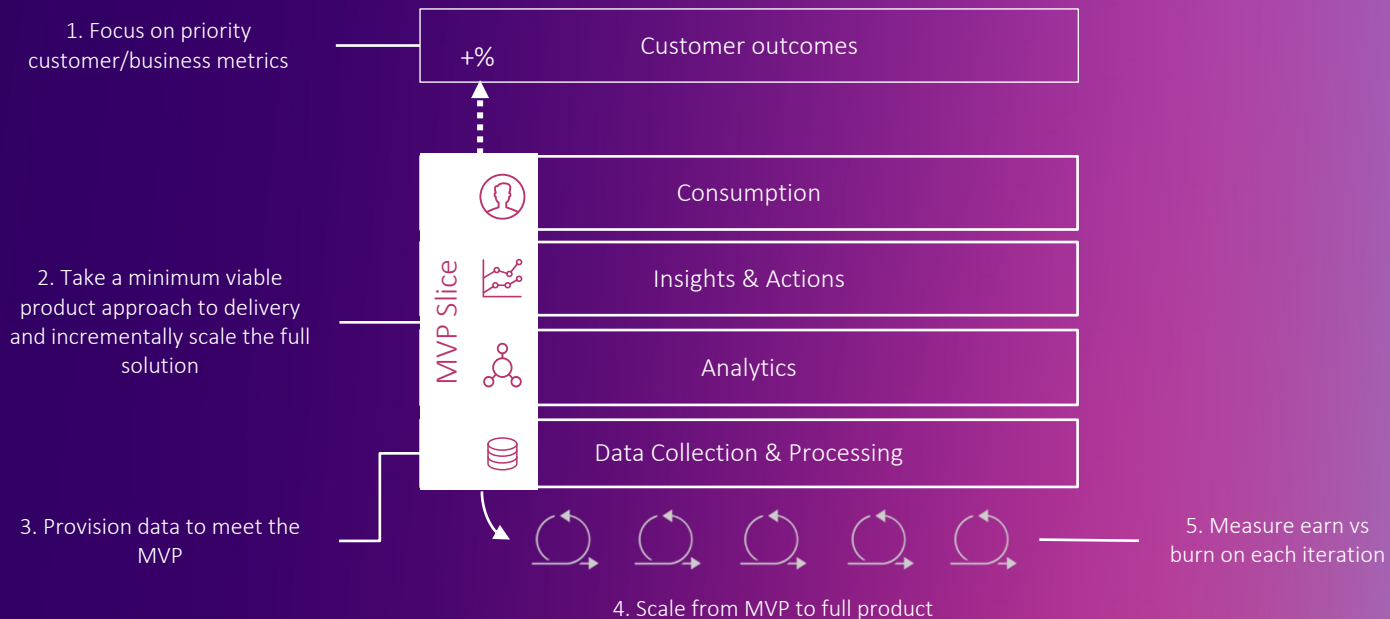Provisioning of data by platform team

aws

Bring the work to the teams: Minimal Viable Product Slice

A minimal viable product (MVP) sliced approach allows 2 pizza teams to build data products that focuses on delivering a minimum set of features that can be released quickly and iteratively to test the viability and potential value of the product. This approach involves breaking down the development process into small, manageable pieces, or "slices," that can be delivered incrementally and built upon over time. Before starting to build a data product, it's important to clearly define the vision and goals of the product. This includes identifying and quantifying the business problem the product is designed to solve, and the customer this is being built for. At **Amazon** we refer to this process as Working backwards, because we work backwards from the customer to the optimal solution to meet their needs, rather than building a solution and then trying to figure out how to use it.

Once the product vision and goals have been established, the next step is to identify the minimum set of features required to deliver value to the user. This may involve prioritizing the most important features, or breaking down larger features into smaller, more manageable slices. With the minimum viable features identified, the next step is to build the first slice of the product. This should be a small, self-contained feature that can be built and released quickly, typically in a matter of weeks. Once the first slice has been built, it's important to test it with users and gather feedback. This feedback can then be used to make improvements and iterate on the product, creating a flywheel effect that iteratively scales the product with each subsequent slice building on the previous one. With each subsequent slice, the product can be built out further and more features and functionalities can be added based on user feedback and demand.

The MVP sliced approach to building data products allows 2 pizza teams to deliver value quickly and iteratively, while also minimizing the risk of investing too much time and resources in a product that may not meet the needs of users or the market. By breaking down development into small, manageable slices, teams can build products that are more agile and responsive to user needs and market trends.

1. Focus on priority customer/business metrics

+% Customer outcomes

MVP Slice

Consumption

Insights & Actions

Analytics

Data Collection & Processing

2. Take a minimum viable product approach to delivery and incrementally scale the full solution

3. Provision data to meet the MVP

4. Scale from MVP to full product

5. Measure earn vs burn on each iteration

aws

11

# Reduce Risk and Move Fast: Decentralize Funding

For 2 Pizza data teams to be empowered to innovate fast requires providing them with autonomy in how they fund and allocate budget to meet with demand for work.

Decentralizing data funding can help organizations become more agile, while also improving the quality and effectiveness of data initiatives. A research paper by IDG finds that 78% of CIOs agree that decentralized ownership of IT spend is good for the future of their business.

By allowing different departments to invest in their own data initiatives, organizations can better align data investments with business objectives, promote ownership and accountability, and optimize the use of resources. This is particularly important in teams that are in functions accountable for creating business differentiation where they need to move fast, and innovate constantly to drive out value for customers.

Decentralizing data funding also results in improved data quality, as departments are incentivized to invest in data initiatives that are most relevant to their specific needs and use cases. This can help ensure that data is more accurate, complete, and timely, which in turn can lead to better decision-making.

Decentralization also fosters a culture of ownership and accountability around data. This helps promote a data-driven culture and encourages collaboration and innovation across the organization.
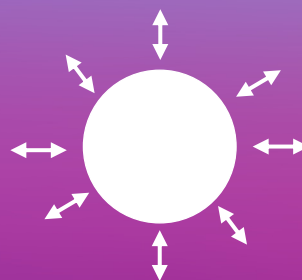
Centralized funding

- ✓ High control of cost
- ✓ High transparency of spend
- ✓ Simple governance
- ✓ Increased standardisation
- x Low autonomy for business innovation
- x Low agility
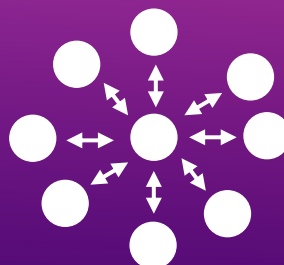
Decentralized funding

- ✓ High agility
- ✓ High business innovation
- - Requires automation for spend transparency
- - Requires federated governance to manage standards and security
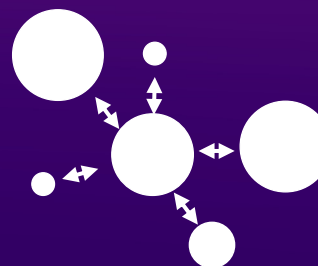
Hybrid funding

- ✓ Manage foundational capabilities centrally
- ✓ decentralised budget for customer facing/ high differentiating capabilities

Reduce Risk and Move Fast: Manage controllable inputs, measure value

Managing controllable inputs instead of outputs allow data teams to be effective by focusing on the things that the team can control, rather than on the outcomes that they want to achieve. Data teams that work backwards from customer needs are able to pinpo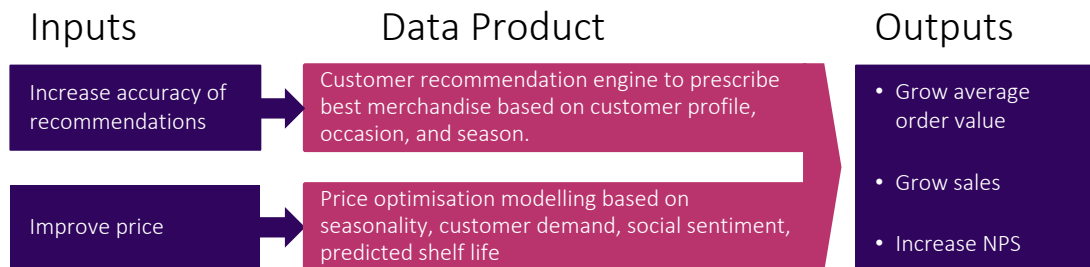int the desired outcomes they want to deliver for customers, and based on this can identify the the inputs they can influence to improve the desired outcome metrics.

For example, a 2 pizza team working in the eCommerce business unit for a fashion retailer is accountable for developing data products to increase customer sales. This team has output metrics to measure Average Order Value, Sales Growth, and customer NPS. In order to improve these metrics the team works backwards from the customer to identify their wants and needs, and map these to the controllable input metrics they can influence in a data use case. In this example fashion customers require improved access to merchandise that matches their style, and affordability (price point). By focusing on the right input metrics (accurate merchandise recommendations, and optimised pricing) the data team can build a data use case that is optimised to drive the right outcomes.

Example: Increasing customer sales and satisfaction in a retail organization

Inputs

Increase accuracy of recommendations	→
Improve price	→

Data Product

Customer recommendation engine to prescribe best merchandise based on customer profile, occasion, and season.

Price optimisation modelling based on seasonality, customer demand, social sentiment, predicted shelf life

Outputs

- Grow average order value
- Grow sales
- Increase NPS

How to Begin

Establishing a Modern Data Community enables data to be accessible and actionable to all as a product, allowing for transparency, speed to action, and collaboration across the business. Doing so requires ongoing close collaboration with the business to shift mindset toward increased ownership of their data, grow maturity in the business to take on greater accountability, and development of technology to allow for autonomy.

Realising this vision requires that there is frequent delivery of value to the business to build enthusiasm and drive the need for more data and information, which further enables insights, and defines new data products to be delivered that scale across the enterprise. To kick start change and gain momentum organisations need to take a flywheel approach. Focus on functions in the business where there are priority opportunities or challenges where data can add value, where there is appetite to do more with data, and there is strong sponsorship from a senior executive in the business.

Specialist data talent is often in short demand across the organisation, and the first 2 pizza data teams might need to begin with being supplemented by centralized data resource, or by leveraging pockets of high talent distributed across the business. Over time autonomous teams will increasingly embed into the business as skills and maturity levels grow. To drive increased scale employ Machine Learning business native tooling where specialist skills are not in place.

Take a Minimal Viable Product sliced approach to delivering on business priorities. Establish the input and output metrics to track value by, and iterate frequently, gathering feedback on not just the value the use case is adding to the business, but also gather feedback on how teaming is working, the need for further data literacy training in the business, and how to continue to improve on the technology solutions being delivered for producers and consumers. Use this feedback to iterate and scale to the wider Think Big vision for a Modern Data Community.

How AWS can help: Data Driven Everything

Next flywheel turn

Transform

Deliver new business value

Mindset

Align business and tech senior leaders on culture, business priorities and financial drivers for the data strategy

People & Process

Focus on building the right organization and process model to support a modern data strategy

Mobilize

Accelerate launching a production-ready value-based use case in the cloud

aws

About AWS

For over 15 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud offering. AWS has been continually expanding its services to support virtually any cloud workload, and it now has more than 200 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media, and application development, deployment, and management from 81 Availability Zones within 25 geographic regions, with announced plans for 21 more Availability Zones and seven more AWS Regions in Australia, India, Indonesia, Israel, Spain, Switzerland, and the United Arab Emirates. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies— trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.com.

About the Contributors

Craig Suckling is the AWS head of Data Strategy, helping customers to reinvent their businesses with data, analytics, and machine learning. Craig has 20 years of experience working as a Chief Data Officer, management consultant, and Startup founder, leading organizations through transformational strategies, navigating change, and generating sustainable growth. He featured in the top 100 data leaders in the UK in 2020, 2021 and 2023.

Hara Gavriliadi is a Senior Strategist at AWS Professional Services helping customers reimaging and transforming their customer experience using data, analytics, and machine learning. Hara has 13 years of experience in supporting organisations to be more data-driven, and turning analytics and insights into commercial advice to enable growth and innovation. Hara is passionate about ID&E and she is an AWS GetIT Ambassador inspiring young students to consider a future in STEM.

Dwayne Browne is a Principal Analytics Platform Specialist in AWS Data Strategy. Through his blend of analytics and design thinking expertise, Dwayne helps organisations innovate and transform their business with solutions that deliver best in class customer outcomes. Dwayne has 18 years of consulting and industry experience across Europe, the US and the Caribbean region as a leader within large organisations and a start-up founder, delivering experience and technology transformation programs in Media, Retail and Telecommunication industries.