The Data-Driven Organization

By Steve Cooper (cpstep@amazon.com), Amy McNee (amymcn@amazon.com), and Jason Berkowitz (jberkowi@amazon.com).

At AWS, we are regularly asked about what a data-driven organization is, what business outcomes it accelerates, the operating model required to sustain data-driven innovation, the role of supporting data strategy, how governance can be enforced in organizations to foster innovation and experimentation, and what technology tenets are required in order to create data architectures that meet ever evolving business needs. The purpose of this whitepaper is to describe best practices and examples based on our experience with companies who are becoming data-driven. The intention is to inspire the end vision, business outcomes, and key tenets in accelerating and scaling the journey.

Introduction

Organizations are overwhelmingly recognizing that data is a key enabler of sustained innovation. Forbes recently published an article that said, "by making 10% more data accessible, a typical Fortune 1000 company will see a $65 million increase in net income."[i] AWS customers are using data to unlock new revenue streams, reduce costs, and mitigate risks. Data is being used to create new customer experiences rather than just report on existing ones. **Engie**, a multinational electric utility company, has put data at the heart of their zero-carbon transition, providing insights on how much carbon dioxide is created, where it is created, and correlation to environmental factors. **FORMULA 1** has redefined the fan experience through real-time insights on driver performance.

At AWS, we call these companies "data-driven organizations" which is defined as:

> *"An organization that harnesses data as an asset, to drive sustained innovation and create actionable insights to supercharge the customer experience"*

Not all organizations have been able to make the transformation. While 99% of blue-chip companies are investing in data initiatives, only 24% have successfully created a data-driven organization. 92% of companies cite culture – people, process, organization, and change management – as the biggest impediment[ii]. Common challenges we hear from companies include understanding how being data-driven can help deliver outcomes, how to identify and prioritize new initiatives, how to build compelling business cases, how to define and shape an integrated business strategy, how to evolve skills and technology foundations, and how to ensure compliance with data privacy, governance, and security. Successful organizations understand that the transformation to becoming data-driven requires more than just technology, it requires a combination of *mindset, people, process, and technology* in order to build and sustain momentum.

Mindset – The Data Flywheel

Data-driven companies focus on solving their customers' most compelling challenges and opportunities, not by thinking in terms of "data platforms", but in terms of "data products". They have shifted to an agile mindset where they rapidly ideate, build, release, and leverage customer feedback in order to enhance the product with additional data. Critically, it is customer feedback and measurable business outcomes that drive further data into the flywheel. By working backwards from the customer and continuously collecting feedback, this ensures they create data products that add value. Built on the right foundations, this allows organizations to think big about what they want to achieve, but start small so they can experiment, learn, and realize value while driving faster innovation. We refer to this cycle of innovation as the "data flywheel"[iii].

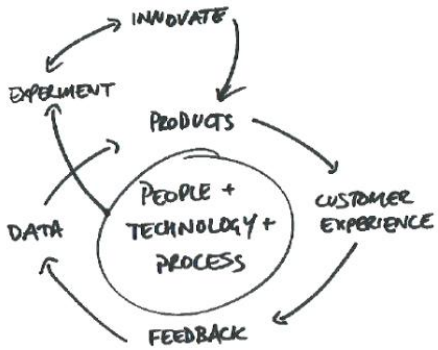

Figure 1: The Data Flywheel

Working backwards from customer challenges ensures the most strategic and diverse datasets are brought together. Data is securely made available to support experimentation and innovation, creating insights that were previously not possible. The lead time to value realization is shortened, allowing companies to focus on developing solutions rather than searching for data and operating platforms. **Discovery Sports Events** is revolutionizing in-door track cycling with a new data driven championship that has data,

analysis, and predictive stats integrated into the fan viewing experience. When Discovery had the idea for a new championship they began with an AWS Data-Driven Everything (D2E) workshop[iv] to work backwards from the customer to prioritize the most compelling features that would create an immersive race experience for the championship launch. The feedback generated by fans viewing and interacting with the race in the velodrome, on screen, and in app continuously informs new features and insights to optimize the viewing experience. This will grow the fan community as the championship scales across their base of 130 million viewers as well as unlocking new customer segments.

The iterative nature of the cycle also ensures that technology adapts to business priorities. The architecture is not fully built in advance, instead technology leaders establish rapid foundational architecture and guiding principles or "tenets" that ensure their architecture scales based upon ever evolving needs. Cloud is an enabler for this, with studies showing a 37.1% reduction in the time-to-market for new products and services[v]. **Ocado**, the world's largest online-only grocery retailer, moved to the cloud in order to experiment faster, deploy faster, and remove undifferentiated heavy lifting in order to focus on the areas where they could add value. The Ocado Smart Platform collects data to train machine learning on diverse use cases including forecasting orders, controlling warehouses in real-time, and routing vans. Engineers can go from conceiving of a basic service to having it deployed to production in under an hour.

People – The Modern Data Community

Mindset is the first step in enabling a data-driven organization. Applying old organizational models to modern, cloud technology is not necessary and won't allow companies to achieve the kind of results and agility they expect. We need an operating model that enables agility and responsibility at the edges, connected by a data marketplace or a data mesh[vi] to create a 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 monolith 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. It is a community of dedicated, autonomous, and small ("two-pizza"[vii]) teams, with a missionary sense of purpose, which, through continuous improvement, results in the team delivering at an ever faster clip[viii]. The modern data community is comprised of three personas; (1) producers, (2) platform, and (3) consumers.

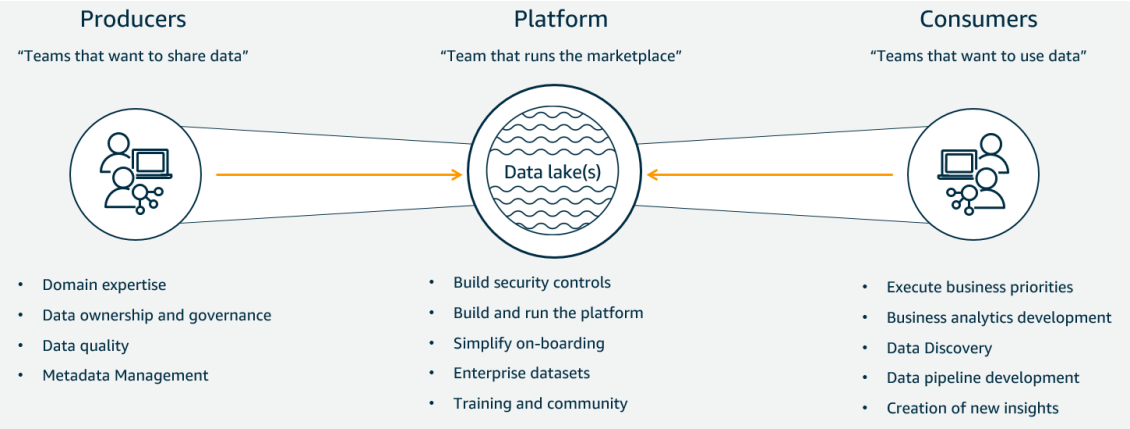

Figure 2: The Modern Data Community

Data Producers are those teams that want to share data. They are domain experts and are typically aligned to upstream business and application teams. Decisions of data stewardship, product ownership, and governance largely reside with these teams. They have the deepest understanding of data quality and business metadata. Our own experience within **Amazon.com** shows the benefits of pushing this responsibility to producers of data. Producers with high quality and reliable data are increasingly leveraged, and this popularity drives positive feedback loops, attracting further usage. High quality and discoverable data drives experimentation, where data is combined in ways previously not considered by producers. This creates a powerful benefit where data becomes the "connective tissue" across internal functions, creating new intra-domain models rather than being tied to organizational silos. We call this connective tissue the "data contract"; the agreement between data producer and data consumer. This contract ensures that consumers' expectations are properly set for things like data quality, frequency, business use and end-of-life SLA's. It is against the data contract that ingestion performance is compared for SLA compliance. The contract is immutable, which means that once a producer commits to a contract, changes can only be made by establishing a new contract. Traditionally, the creation and maintenance of meaningful data contracts has been complex, however with machine learning techniques, the effort is greatly reduced. Profiling enables data stewards to set data quality rules, personally identifiable information (PII) can now be identified at

scale, and classifications can aid in the tagging of data for easy discovery. Entitlements can be managed at scale, providing robust security and audit capabilities.

As experimentation and innovation occurs, the producer's data is further monetized, realizing value for the organization and the producer. Customers like **BMW** have connected their highly secure data backend, made up of 188 data producers across 15 business domains, to an open data catalog, this supports innovation of multiple use cases including vehicle health indicators to issue resolution.

Data Platform is the team that is responsible for the operation of the "data marketplace". This is the technology that enables the data community. At **Amazon.com**, since 2017, we have deliberately used the word marketplace[ix]. We believe that data is something that has intrinsic value and the most successful marketplaces are those that have a range of diverse products, are reliable, and provide a delightful producer and consumer experience. Platform teams are typically small, leveraging the power of the cloud to provide simple on-boarding of new data, ease of browsing, comparison of products, reliability, durability, scalability, highly secure transactions, and ease of purchase. 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. Customers such as **Netflix** have demonstrated the power of strong platforms, managing 700+ billion events per day, supporting a culture of data discovery, and open sourcing engineering solutions to the broader community.

Data Consumers are the individuals, teams, and machines that want to use data. This is a diverse group, including business analysts, data engineers, software engineers, and data scientists. Similar to the data producers, consumers are not aligned into monolith IT organizations, they are decentralized and aligned into their business domain. Their proximity to the business ensures they understand the business priorities and are able to test business hypotheses. Data consumers want to minimize non-value adding time spent searching for data, requesting access, and testing quality. Data consumers want - at speed - to be able to search, discover, understand quality, trust, subscribe, and run analysis in their tool(s) of choice. **BBVA** is an example of a customer who has run data literacy and data guild certification at scale, dramatically increasing analyst data and analytics literacy in order to drive the benefits of decentralized innovation.

Once the data becomes accessible to a diverse group of data consumers, you need to focus on driving data literacy across the organization. You can't become data-driven if the ability to understand and interpret data is not broadly available across the enterprise. Historically, companies have relied on Centers of Excellence, business catalogs, and other documentation to create data literacy programs. However, in many organizations, the ability to interpret data still lies with only a few analysts who demystify charts, graphs, and tables so they can be understood by less expert stakeholders. In the past few years, technology has evolved to create analytics that are already in a "narrative" format to ensure clarity on what the data is indicating. AI/ML has been integrated into traditional Business Intelligence capabilities to automatically detect anomalies and improve the ability for less technical stakeholders to find the most important information. In addition, new methods for querying data using natural language makes it feasible for anyone to ask the next most important question without having to rely on those people in the organization who are proficient in SQL. **Panasonic Avionics Corporation,** the world's leading supplier of in-flight entertainment and communication systems, collects large amounts of anonymized data that help optimize the experience for both our airline partners and their passengers. Data consumer teams are able to make use of machine learning-powered anomaly detection and Amazon QuickSight Q in order to interpret the business context of natural language questions, provides synonyms, and provide an answer with no complex interpretation needed.

<u>Process – Invent and Simplify a Balance between Governance and Agility</u>

With mindset and data community established, data-driven organizations have reinvented their data processes in order to ensure they balance the needs of evolving governance requirements while fostering agility and innovation. Successful companies are taking advantage of step changes in technology in order to sustain the change to processes.

In the vision of the modern data community, we intentionally align data responsibilities to the data producers. This is a departure from many traditional operating models where the role of data governance and stewardship commonly resided within a central organization (often IT). For too long, the centralization of this activity was aligned with a view that data governance was hard; it was a blocker to innovation. The decentralization of responsibility is both mandatory in order to enable agility and a recognition of the power of data governance to sustain innovation. This change often means a transfer of power from centralized teams to the "real" business owners who have a greater vested interest in the value created.

In order to enable this transfer of responsibility, we need to ensure that data governance, security, and privacy are not compromised. We leverage advances in automation and machine learning in order to both comply with governance standards as well as raising the bar on data quality, governance, and compliance. Automation is also key to providing enhanced services to data producers. This reduces institutional resistance from upstream stewards, ensuring that we are improving their day-to-day jobs.

Customers such as **Cox Automotive**, working with Collibra, an AWS ISV partner and leader in data governance, connect traditional data producers, data engineering teams, and data scientists via a data catalog in order to identify the most strategic data assets and ensure the appropriate level of data governance. Within **Amazon.com**, our data engineering teams created Deequ[x], an open-source tool that allows data producers to calculate data quality metrics, define and verify data quality constraints, and be informed about changes in the data distribution. Instead of manually creating checks, Deequ suggests rules and is designed to scale to large datasets. Amazing things happen when you give talented teams the autonomy and leverage to invent and simplify. The decentralization of the data community is enabled through automation of observability and auditing.

Decentralization does not imply a lack of structure. A key aspect of data governance and quality is the ability to comply with structured taxonomies and ontologies. Within **Amazon.com**, we wrap taxonomies in APIs in order to ensure that teams are able to consume "taxonomy as a service". Advances in technology have helped companies automate knowledge graphs in order to structure and organize information. **Siemens Energy AG** is structured with business units spanning across Central Power Generation, Distributed Power Generation, Industrial Applications, Transmission, New Energy Business, and the Siemens Gamesa Renewable Energy's (SGRE) on and off shore wind businesses. Siemens Energy AG uses Amazon Neptune and partner solutions to represents the hierarchy of product (e.g. gas turbine) configurations. They have moved heterogeneous and manually intensive bills of material into a digital representation of the turbine taxonomy. This allows Siemens to find inconsistencies, curate data, and constantly monitor data quality, enabling the root cause analysis to source and resolution at the origin.

Organizations are facing evolving compliance needs for regulations, data privacy requirements, and guidance on protecting additional types of sensitive data. This can make identifying and protecting sensitive data at scale complex, expensive, and time consuming. Companies leverage Amazon Macie to ensure safe, secure, and responsible innovation. Amazon Macie uses machine learning and pattern matching to discover and protect sensitive data in AWS. Health information technology companies can ingest patient and hospital data and make near-real-time predictions about patient care and hospital operations—such as hospital capacity and length of individual patient stay – that healthcare systems and clinicians can use to make more informed decisions. Data privacy is central to these solutions, allowing researchers to work on de-identified patient health data for research and organizing data into data sets in research-ready formats. HIPAA compliance of AWS services was required in order to adequately safeguard protected health information (PHI).

Technology – The Lake House

AWS Services provide the technology required to sustain flywheel momentum and balance the agility and governance of a scaled modern data community. Companies need to be able to easily spin up new capabilities to support experimentation, reduce the number of times data is copied and moved, ensure that data is protected, and costs are controlled. While managed services in the cloud eliminate many of the challenges surrounding hardware configuration and procurement, data center challenges, and other operational overhead, to fully achieve this, companies cannot use the same architectural approach that was available in the days where relational technology was the only choice. A modern data architecture is required. This is built on tenets that ensure the architecture scales based upon ever evolving needs, for example, purpose-built data services, data governance/security as job zero, and API-centric design.

This "Lake House architecture"[xi] [xii] is comprised of (1) open and scalable data lakes, (2) unified governance, (3) purpose-built data services, (4) seamless data movement, and (5) cost-effective performance. This isn't just about integrating your data lake and your data warehouse. It is about connecting all of your purpose-built services into a coherent whole. The Lake House lets you turn on capabilities as you need them to solve business problems.

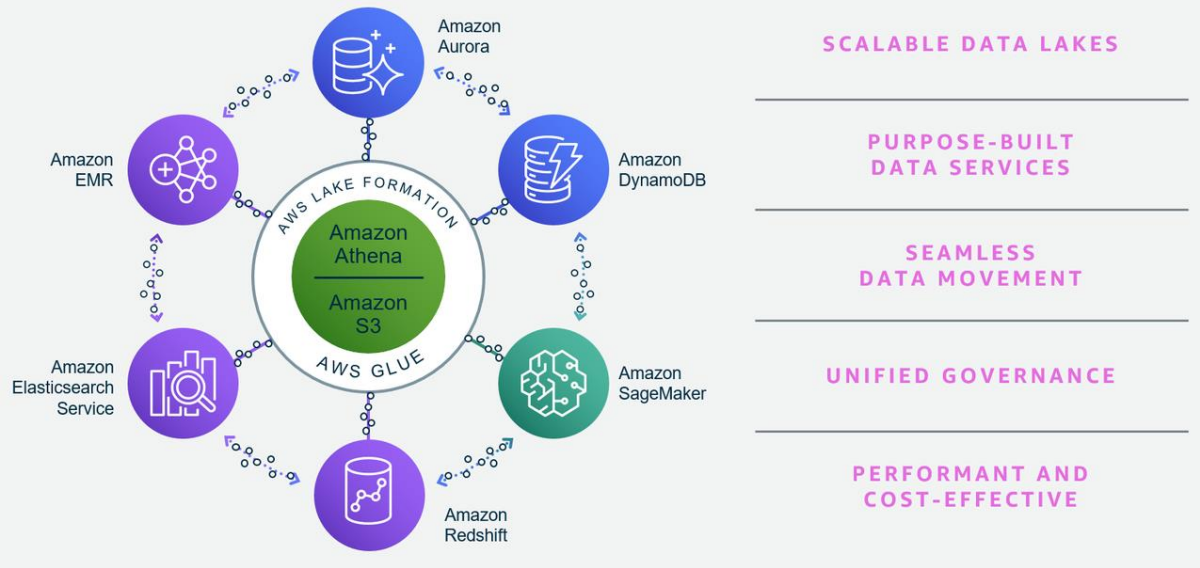

Figure 3: Conceptual Lake House Architecture

Seamless data movement is enabled with self-service ingestion made available to the producers using Amazon Kinesis, AWS Database Migration Service, and AWS Glue. Tooling is standardized via AWS CloudFormation templates in order to ensure adoption by diverse sets of producers. In addition, frameworks like AWS Serverless Data Lake Framework[xiii] (SDLF) provide the ability to run data quality at scale, ensuring that producers get early notification of issues with their data quality or service level agreement. A common pattern is "inside out", where data producers land data directly into Amazon Simple Storage Service (S3) to cost effectively build and scale a data lake of any size in a secure environment which is protected by 99.999999999% (11 9s) of durability. In other use cases, data producers will ingest data into their data lake from the "outside in", for example, landing transactional data in a relational database and bringing portions of this data into the data lake to support full view of the customer profile. Once the data is in S3, AWS services such as Redshift Spectrum, Athena, and EMR allow you to operate on your data directly in the data lake without having to copy or move it. AWS Glue DataBrew, SageMaker, and QuickSight can also leverage data directly in S3 for data wrangling, data visualization, and model development. This reduces the overall cost and time required to enable new use cases. Organizations like **Swimming Australia**, the governing body for swimming in Australia, used SDLF to give coaches a better picture of how athletes are performing. Their data lake is now the location for evidence-based decision making, containing both competition and training data. It supports multiple analysis tools, some specifically built for swimming, and multiple users including coaches, sport scientists, and universities. This supports community innovation, for example, the "H2Grow Program", built by the University of Sydney, which identifies high performing swimmers at a young age.

We have seen how advances in technology have enabled changes to processes; data governance, data quality, and compliance. AWS Services have enabled the move from centralized to decentralized governance. AWS Lake Formation collects and catalogs data from databases and object storage, moves the data into an Amazon S3 data lake, cleans and classifies data using ML algorithms, and secures self-service access to sensitive data. Amazon S3, AWS Lake Formation, and AWS Glue provide the technological foundation of the data marketplace. They provide a single location for data producers to onboard data and for consumers to search, discover, understand quality, subscribe, and run analysis in their tool of choice. AWS Lake Formation is used to share datasets securely between producers and consumers. Data does not need to be copied and ownership remains with the data producer. Many companies choose to update and expose metadata in a single data catalog, leveraging AWS Glue to crawl, extract metadata, and serve as the repository for consumers to search for data. For organizations wanting to go further, they can leverage Amazon Elasticsearch to perform free-text searches at speed and, when combined with Amazon Neptune, the power of graph means that data consumers are recommended data sets. **BMW**, leverages a combination of Amazon S3 for their data lake, AWS Glue, and a custom frontend unified data portal and API layer to expose over a petabyte of data from more than 200 producers in 25 decentralized locations across both on-premise and cloud databases to more than 5,000 business analysts and data scientists creating insights. The portal allows business analysts to ingest data and perform analysis as a seamless experience.

AWS allows for a seamless integration of data sets to the consumers' analytics tool of choice through common access with enforced access controls and audit tracking. Frequently termed the "last mile", this ensures that when consumers find their dataset(s), they are able to use their preferred analytics tools in order to perform analysis. AWS offers the broadest and deepest set of purpose-built analytics, AI, and machine learning services. This range of services enables analysts with everything from ad-hoc data query at scale (Amazon Athena), transformation (AWS Glue), data warehouse dashboards (Amazon Redshift, Amazon QuickSight), machine learning (Amazon SageMaker), and AI services (such as text, fraud, and recommendations). Innovations in Amazon QuickSight and

AWS Glue DataBrew enhance data literacy in organizations, using integrated AI/ML and natural language query so that business users can get the information they need in an understandable way.

The Outcome

Applying a modern methodology for mindset, people, and process is as important as modernizing technology to create a data-driven organization. This enables and fosters the innovation of business outcomes through data. Data-driven organizations are using data to unlock new revenue streams, reduce costs, mitigate risks, and enhance their customer experience. Recent research by IDC[xiv] highlighted a 9-month payback period on investment and 415% five-year ROI. Customers such as Netflix, Formula One, Engie and many more have made this transformation with AWS and the pace is accelerating[xv].

AWS offers many programs to help you get started and scale your data-driven organization. AWS Data-Driven Everything (D2E) helps jumpstart your data flywheel and AWS Data Lab[xvi] helps build initial solutions. “The AWS D2E program was instrumental in driving alignment between our business and technology leaders to paint a vision of what we ultimately want our most important stakeholder, the injured worker, to experience when interacting with Bardavon. D2E not only provided a well thought-out long-term vision using the working backwards methodology, but also recommended a fast, practical way for us to take our first steps in the journey leveraging data and analytics in the cloud.” Matt Condon, CEO, Bardavon Health Innovations

For more information, please contact your AWS Account Executive.

About the Authors

Steve Cooper, Head of Worldwide D2E Program, AWS.

Steve is the worldwide lead for Data-Driven Everything (D2E) at Amazon Web Services. D2E provides a use-case driven framework to help customers create a compelling data-driven vision, accelerate the flywheel for their most challenging use cases, build experience, and scale with a “go-big” plan and 6 to 9-month roadmap of priority projects. Prior to joining AWS, Steve led Accenture's big data platform and AI/ML studio services, driving Accenture's transformation to an intelligent enterprise.

Amy McNee, Head of Worldwide Analytics Platform Strategy, AWS.

Amy leads a worldwide team helping customers with their analytic strategy including both technology and the people/process to achieve specific business outcomes driven by data. Amy has over 25 years of experience covering the end to end ecosystem for analytics and business intelligence with companies including SAP, Workday and AWS.

Jason Berkowitz, Head of Worldwide AWS Professional Services Data Analytics, AWS.

Jason is the worldwide leader of AWS Professional Services Data Analytics best practices, methodologies and analytics solutions. He comes from a background in Machine Learning, Data Lake Architectures and helping customers become data-driven. He is currently working helping customers shape their data lakes and analytic journeys on AWS within Professional Services.

Prior to joining AWS, Jason spent 20+ years in analytics architecture, development and leadership roles at companies like CME Group and Accenture.

Appendix

Customer References

- Amazon: re:Invent 2019 and case study
- BBVA: https://www.bbva.com/en/bbva-wants-to-turn-its-employees-into-digital-ninjas/
- BMW: https://aws.amazon.com/solutions/case-studies/bmw-group-case-study/ and re:Invent 2020
- Cox Automotive: https://www.youtube.com/watch?v=8V5W-QnlqkQ&feature=youtu.be
- Discovery Sports Events: Video and Press Release
- Engie: case study
- FORMULA 1: https://aws.amazon.com/f1/
- Netflix: https://aws.amazon.com/solutions/case-studies/netflix/
- Ocado: https://www.youtube.com/watch?v=Bp82g_TuYyg and https://www.youtube.com/watch?v=g2uMfpAHdGY
- Panasonic: https://aws.amazon.com/blogs/big-data/bringing-the-power-of-embedded-analytics-to-your-apps-and-services-with-amazon-quicksight/
- Siemens Energy: https://d1.awsstatic.com/AWS%20Databases/metaphacts-Case-Study-Turbine-Spare-Parts-Management-at-Siemens-Energy.pdf
- Swimming Australia: case study

Endnotes and Technical References

---

[i] Dykes, Brent (2019). *The Four Key Pillars to Fostering a Data-Driven Culture.* Available online at: https://www.forbes.com/sites/brentdykes/2019/03/28/the-four-key-pillars-to-fostering-a-data-driven-culture/

[ii] New Vantage Partners LLC (2021). *Big Data and AI Executive Survey 2021.* Available online at: https://www.businesswire.com/news/home/20210104005022/en/NewVantage-Partners-Releases-2021-Big-Data-and-AI-Executive-Survey

[iii] Collins, Jim (2001). *Good to Great.* Random House Business; 1st edition.

[iv] AWS Data-Driven Everything. Contact your account manager.

[v] AWS & Omnicom Group (2020). *Cloud Value Benchmarking Study Quantifies the Benefits of Cloud Adoption*. Available at: https://pages.awscloud.com/rs/112-TZM-766/images/cloud-value-benchmarking-study-quantifies-cloud-adoption-benefits.pdf

[vi] Dehghani, Zhamak (2019). *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh.* Available at: https://martinfowler.com/articles/data-monolith-to-mesh.html

[vii] Slater, Dan (2020). *Elements of Amazon's Day 1 Culture.* Available at: https://aws.amazon.com/executive-insights/content/how-amazon-defines-and-operationalizes-a-day-1-culture/

[viii] Cagan, Marty (2018). *Inspired. How to create tech products customers love.* John Wiley & Sons; 2nd edition.

[ix] AWS (2018). *AWS re:Invent 2018: How Amazon Uses AWS Services for Analytics at a Massive Scale (ANT206).* Available online at: https://www.youtube.com/watch?v=PitJL9vOotc

[x] Serverless Data Quality with Deequ on Glue: https://github.com/aws-samples/amazon-deequ-glue

[xi] Pathak, Rahul (2020). *Harness the power of your data with AWS Analytics.* Available at: https://aws.amazon.com/blogs/big-data/harness-the-power-of-your-data-with-aws-analytics/

[xii] Shankar, Nivas (2021). *Design a data mesh architecture using AWS Lake Formation and AWS Glue*. Available at: https://aws.amazon.com/blogs/big-data/design-a-data-mesh-architecture-using-aws-lake-formation-and-aws-glue/

[xiii] Serverless Data Lake Framework (SDLF): https://github.com/awslabs/aws-serverless-data-lake-framework

[xiv] IDC (2020). *The Business Value of AWS Data Lakes, Analytics, and ML Services.* IDC. Available at: https://aws.amazon.com/resources/analyst-reports/?analyst-reports-main.sort-by=item.additionalFields.datePublished&analyst-reports-main.sort-order=desc&awsf.analyst-reports-category=categories%23big-data

[xv] New Vantage Partners LLC (2021). *Big Data and AI Executive Survey 2021.* Available online at: https://www.businesswire.com/news/home/20210104005022/en/NewVantage-Partners-Releases-2021-Big-Data-and-AI-Executive-Survey

[xvi] AWS Data Lab. https://aws.amazon.com/aws-data-lab/