The State of Data Today

Learn about how third-party data is empowering organizations to transform the way they operate

First published May 10, 2023

Last updated May 10, 2023

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Introduction

The great data race continues unabated – and the most forward-thinking and progressive organizations are looking to win by providing fast, reliable access to data anytime, anywhere, throughout their enterprise. What does this fast, reliable, universal access to data look like in business terms? Look towards the bottom line.

Organizations are only as good as the data they possess, but that only tells part of the story when it comes to first-party data. Forrester[1] found that data-driven organizations are 162% more likely to significantly surpass revenue goals than laggards. Companies use third-party data to fill gaps and create personas, but they can also add another dimension to their analytics, strategies, and bottom lines. MIT Sloan Management Review[2] found that 'the most analytically mature organizations use more data sources.' Their study found that 'analytical innovators' were four times more likely than less mature companies to use data from customers, vendors, regulators – and competitors. Yet reliability and trustworthiness is key: as more data sources come in, questions get asked around what the data means, who owns it, and where it came from.

In this paper, we examine and analyze the state of data today across the financial services, healthcare and life sciences, media and entertainment, and consumer packaged goods (CPG) and retail industries, including examples of how organizations are procuring third-party data and are optimizing the way they operate. We will specifically go over how you can transform your business and get started with AWS Data Exchange, where you can easily find, subscribe to, and use third-party data in the cloud to generate your own insights. This paper is primarily aimed at a business audience; C-suite and senior management working in IT, Finance, or Business Intelligence functions.

Note: the term 'data subscriber' and 'subscriber' as used in this whitepaper refers to the buyer i.e. the user of a third-party data catalog, while 'data provider' and 'provider' refers to third-party data companies listing their data in a data catalog.

[1] "How a Data Catalog Can Help Your Business Reach New Heights", Database Trends and Applications, https://www.dbta.com/Editorial/Trends-and-Applications/How-a-Data-Catalog-Can-Help-Your-Business-Reach-New-Heights-144574.aspx

[2] "Using Analytics to Improve Customer Engagement", MIT Sloan Management Review, https://sloanreview.mit.edu/projects/using-analytics-to-improve-customer-engagement/

Chapter 1: The State of Data Today

If you look up the state of the big data landscape today, words like ‘tsunami’ and ‘deluge’ often appear – with good reason. According to an analysis from Domo[3], there are 2.5 quintillion bytes of data created each day, with 90% of the world’s data generated in the last two years alone. This correlates to 912.5 exabytes in a year, or 0.912 zettabytes. By comparison, in 2012 the world's all-time data output exceeded one zettabyte for the first time[4].

These statistics are a good indicator of how much data is flying around, but they do not scratch the surface on how and why the strongest organizations are in this position. The truth is that while all organizations are utilizing data, comparatively few are doing so as effectively as they can. Gartner[5] predicted at the beginning of 2019 that by 2022, only 20% of organizations’ analytic insights will deliver business outcomes. Two years later, Kearney[6], in its Analytics Impact Index, offered a figure which suggested little had changed: fewer than one in five organizations were hitting all of the company’s benchmarks required to be data leaders in today’s environment.

There are several factors which separate the best from the rest, some of which are interlinked. These include getting the right technology, getting cultural buy-in, and using third-party data effectively.

On the tech side, data leaders are moving beyond just platform adoption to embracing future-proof tools and processes, from data science and analytics, to management, to architecture and engineering, to data and information governance. Another key example is agility and automation; data leaders are leveraging their

[3] “How Much Data Do We Create Every Day?”, Bernard Marr, https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

[4] “Mobile Networks in a Zettabyte World: Trends from Cisco’s Visual Networking Index”, GSMA, https://www.gsma.com/spectrum/wp-content/uploads/2012/06/Dr_Robert-_Pepper_Cisco_Public_Policy-Forum_Data_Demand.pdf

[5] “Our Top Data and Analytics Predicts for 2019”, Gartner, https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

[6] “Leader or laggard? It’s all down to the data dividend,” Kearney, https://www.kearney.com/digital/article/-/insights/leader-or-laggard-its-all-down-to-the-data-dividend

platforms to translate data into immediate insights and actions and automating many decisions in real-time.

Organizations' data architectures are changing too. The need for third-party data across the organization can lead to data silos; therefore the need for a connective element, or single data architecture, is apparent. One emerging example of this is the data lakehouse[7]. Whereas a data warehouse is specifically designed for data analytics, a data lake is a centralized repository for all data, from structured, to semi-structured, to unstructured. A lakehouse architectural approach looks to take the best elements of data warehouses and data lakes – small, coordinated data versus massive, uncoordinated data respectively – and merge them, providing improved controls and tools.

Amazon Web Services (AWS) first described the 'lake house architecture' approach at the end of 2020[8], involving several products, including Amazon SageMaker, Amazon Redshift, Amazon DynamoDB, AWS Lake Formation, and AWS Glue, which enables organizations to discover, prepare and integrate all their data at any scale. The architectural approach "acknowledges the idea that taking a one-size-fits-all approach to analytics eventually leads to compromises. It is not simply about integrating a data lake with a data warehouse, but rather about integrating a data lake, a data warehouse, and purpose-built stores and enabling unified governance and easy data movement."

From a cultural perspective, there are several key tenets which define a truly data-driven organization. The first tenet is to be inclusive. If you want an authentic data-driven organization, then data has to be used at every level of the organization. The business only moves forward if all stakeholders are keyed in. There is a misconception around data being purely a technical pursuit; it must be emphasized that data is an integral part of business operations and processes. The next tenet is to be more curious – or rather, less complacent – and moving away from instinct as part of the business decision-making process. This also includes upskilling employees; encouraging discovery with data, and empowering workers across the business to develop their skills and make data-driven decisions. Such a process does not happen overnight – and so the third tenet is to be incremental, in both culture and data. A

[7] "Data Architecture Trends in 2022", Dataversity, https://www.dataversity.net/data-architecture-trends-in-2022/

[8] "Harness the power of your data with AWS analytics", Amazon Web Services, https://aws.amazon.com/blogs/big-data/harness-the-power-of-your-data-with-aws-analytics/

‘big bang’ approach[9] to data will deliver results in the short term, but the needs of the business – and the questions which need to be answered by data and analytics – change over time.

The right culture and the right technology are often linked. The right technology procurement decisions are of course extremely important, but the strategy needs to drive the technology, not the other way around. “Everyone loses when business strategy and data strategy are not aligned,” writes Emanuel Younanzadeh in Forbes[10]. Tying all the threads together and having the right culture turns a good company into a great one.

Including third-party data in this strategy, for reasons of adding an extra dimension to analytics, strategies and bottom lines as explained earlier, can be seen as the last piece of the jigsaw – and again, it is inextricably entwined with technology strategy.

There are myriad benefits of third-party data across industries. The Data & Marketing Association (DMA)[11] outlines the key ones: adding flavor and depth to first-party data; giving greater context to why customers behave in particular ways, such as market events; or understanding existing customer characteristics to find more like them in different audiences. The latter is certainly among the most important, in terms of ‘unlocking the customer’ and going ‘beyond the transaction.’ “Adding this richness to what you already know unlocks a whole host of uses beyond your brand’s owned boundary walls,” the DMA notes.

Organizations are now understanding what it takes to deliver on the promise that a good data strategy – be it around first-party or third-party data – brings, and the architectural and cultural changes required. It can be a slog; it may involve significant change to how you and your organization does business; and it won’t happen overnight. But ignoring it is simply not an option. In the next chapter, we will explore data strategies by industry, and understanding what good looks like across financial services, media and entertainment, healthcare and life sciences, and retail and CPG.

[9] “The Core Tenets of Building a Data-Driven Organization,” Spaulding Ridge, https://www.spauldingridge.com/articles/the-core-tenets-of-building-a-data-driven-organization/

[10] “Are You Really a Data-Driven Organization?”, Forbes, https://www.forbes.com/sites/forbescommunicationscouncil/2021/12/29/are-you-really-a-data-driven-organization/?sh=2515efa43906

[11] “Unlocking the Customer with Third-Party Data – what do you need to consider?”, Data & Marketing Association, https://dma.org.uk/article/unlocking-the-customer-with-third-party-data-what-do-you-need-to-consider

Chapter 2: Analyzing Data Strategies by Industry

We are seeing data-sharing use cases – and in particular cases, success stories – proliferate as more organizations begin pursuing opportunities to expand their data assets. Here are a few brief examples.

Financial services

The financial services industry naturally runs on data. This can be visualized in a couple of ways. A 2022 analysis from TRG Screen[12] found that spending on market data in the financial services industry rose 6% year-on-year, to $33 billion. Almost three quarters (72%) of UK financial services firms surveyed by the Bank of England[13] in the same year reported using or developing machine learning applications, up five percentage points from 2019, with enhanced data and analytics capabilities the primary benefit. In the banking sector, this data can include analytics around clients' trading activities, as well as using customer data points, such as credit scores and financial transactions, to more accurately forecast risk. For the insurance sector, this can include analytics to build risk profiles of potential applicants, as well as utilizing various customer data points, such as telematics or smartphone apps, to offer more personalized or targeted offerings.

Key business benefits of a strong data analytics strategy in financial services include the creation of more secure processes; increasing operational efficiency by making faster, better decisions; and maintaining competitive advantage. Real-time data analytics can enhance organizations' competitive advantage by responding to trends quickly; risk and fraud can be better prevented through machine learning algorithms; and customers can be better targeted and segmented to improve their product marketing.

[12] "2022: A transformative year in financial market data?", Finextra, https://www.finextra.com/blogposting/21857/2022-a-transformative-year-in-financial-market-data

[13] "Machine learning in UK financial services", Bank of England, https://www.bankofengland.co.uk/report/2022/machine-learning-in-uk-financial-services

Third-party data can enhance how financial services organizations operate. In a recent AWS webinar[14] Akanksha Sharma, Director, Cloud Integration, Analytics and Data Science at Citi – a longstanding AWS customer – explains how the company combines third-party data with internally-generated information in order to improve overall algorithms. "Of course, there is a lot of interest in understanding how location data, for example, can be used, especially on the retail side of the banking services, and how we can provide faster, cheaper services to our clients on a day-to-day basis – and as real-time as we can go," said Sharma.

Citibank uses multiple third-party data sources; news updates for scoring particular investment portfolios, ESG scores, and market data for tickets. These can come from channels such as government data sources for interest rates, currencies, and exchange rates. The company regularly deals with quantitative analysts, defined by Investopedia[15] as the 'rocket scientists of Wall Street.' These analysts are highly technically literate, being well-versed in using different data formats and Python-like languages. However, not all users are like this, so work is spent with less technical users in making the data available with the right accuracy, consistency, and completeness. Time is of the essence here: for some use cases, milliseconds of latency are demanded, without compromising on content or data quality. "For any business user, the key is to get access to accurate information in the least time-consuming way," says Sharma.

Media and entertainment

For media and entertainment, data and analytics use cases can include[16] predicting audience interest, engagement and disengagement, as well as optimization and monetization. Third-party data can include metadata across content, alongside historical ratings data. IMDb for example has metadata from over nine million titles, 12 million cast and crew listings, and global box office grossing figures.

Perhaps the most evident example of a data and analytics use case is in recommendation engines. A research paper released by Netflix (Lamkhede, Kofler

[14] "How To Maximize The Capabilities Of Your Data Assets, A Focus On Third-Party Data", Amazon Web Services

[15] "Quants: The Rocket Scientists of Wall Street", Investopedia, https://www.investopedia.com/articles/financialcareers/08/quants-quantitative-analyst.asp

[16] "Big Data in Media and Entertainment Industry", Analytics Steps, https://www.analyticssteps.com/blogs/big-data-media-and-entertainment-industry

2021)[17] outlines the two key requests from users based on qualitative and quantitative data which a recommendation engine answers, 'I know what I want, I need you to get it for me' or 'I don't know what I want, let's understand what you have.' Much like quantitative analysts in financial services, recommendation engines frequently utilize Python[18]. The media and entertainment industry is frequently cited as one of the best opportunities for data scientists, with in-demand jobs including forecasting analyst, content analyst, and business intelligence developer.[19] Another important use case for the industry is catalog matching. Accurate and timely metadata for titles is essential for stakeholders such as streaming service providers. Yet provider metadata is often sparse; perhaps just a title and a run time. Data providers can offer enriched and validated metadata with unique and consistent identifiers, which is vital for building a canonical framework for organizing content catalogs.

Healthcare and life sciences

Key opportunities abound in disease tracking and modeling, diagnosis and prognosis analysis, research, and treatment personalization and optimization. IDC[20] estimated that, through 2020, approximately 270GB of healthcare and life science data was created for every person in the world. The increasing importance of real-world data (RWD) can also be noted as part of third-party data. This includes electronic health records (EHRs) and claims data, which have use cases around developing precise medicine strategies through digital pathology and radiology data, as well as developing a 'deeper understanding of mechanisms of action and clinical outcomes.' The Covid-19 pandemic, and the requirement for urgent collaboration, has shone a

[17] "Recommendations and Results Organization in Netflix Search", S. Lamkhede, C. Kofler, https://arxiv.org/pdf/2105.14134.pdf

[18] "What is a Recommender System in Machine Learning?", Codecademy, https://www.codecademy.com/resources/blog/what-is-recommender-systems-machine-learning/

[19] "What Industry Hires The Most Data Scientists?", Career Karma, https://careerkarma.com/blog/what-industry-hires-the-most-data-scientists/

[20] "The Data Dilemma and Its Impact on AI in Healthcare and Life Sciences", IDC, https://blogs.idc.com/2021/06/23/the-data-dilemma-and-its-impact-on-ai-in-healthcare-and-life-sciences/

light on greater urgency in this field, with use cases proliferating from testing, to data on viral variants, host outcomes, and vaccination status[21].

Retail and consumer packaged goods (CPG)

There are a huge number of opportunities for data and analytics relating to the retail and wider CPG industry. Data can range from customer transaction information and purchase loyalty, to pricing and sales trends, to data which analyzes the in-store position of certain products. Third-party data can include weather, population and footfall data to forecast customer and inventory needs – to the extent of where to position a new store – to market research data analyzing brand sentiment, to targeting customers' purchasing patterns.

Alongside better inventory management, wider use cases include personalizing the customer experience, both in-store and through eCommerce and mCommerce, and creating greater efficiency throughout the supply chain, from sourcing to point of sale.

AWS[22] gathered more than 40 CPG leaders at the 2022 CGT (Consumer Goods Technology) Analytics Unite conference to explore key actions, measurements for success and blockers for various data and analytics use cases. The most popular use cases were optimizing on-shelf availability, using geo data in targeting and personalization, optimizing product assortment and distribution, and democratizing data for self-service analytics. Key observations from the session included the need for cross-collaboration in all use cases, which is difficult as many CPGs have built strong brands while operating in silos. Similarly, there is still work to be done in terms of sharing granular data in near real-time between retailers and CPGs.

The nuances for different industries in terms of their data and analytics strategies can now be seen. Aligning this knowledge with the increasing movement of organizations towards third-party data services, the next chapter will go deeper on understanding these services, as well as more specific examples of how forward-thinking organizations are best utilizing third-party data.

[21] "COVID-19 data sets and APIs on AWS Data Exchange," Amazon Web Services, https://aws.amazon.com/data-exchange/covid-19/

[22] "How CPGs are using data and analytics to drive change and value," Amazon Web Services, https://aws.amazon.com/blogs/industries/how-cpgs-are-using-data-and-analytics-to-drive-change-and-value/

Chapter 3: Understanding Third-Party Data Services

Overview

With buying and selling data on the rise, and the idea of third-party data firmly established, the cloud-based third-party data service, or cloud marketplace, is an obvious next step: almost all (96%) of the 753 respondents in the most recent Flexera State of the Cloud report[23] were using public cloud, with more than half of respondents spending at least $2.4 million on it each year. The Covid-19 pandemic and shift to remote working has accelerated this migration. Canalys[24] noted that, referring to wider cloud marketplaces, 2021 saw $4.1 billion of sales go through these marketplaces, with the figure rising to $25 billion by 2025 at a compound annual growth rate (CAGR) of 59%. This has been fueled by growth in cloud infrastructure generally; 2021 Q2 saw worldwide cloud infrastructure services grow by more than a third, to $47 billion.

A study from Forrester Consulting in March 2022[25], commissioned by AWS, explored organizations' cloud software and/or data procurement processes, polling 725 global decision makers. Overall, four in five (79%) respondents said that in the next 12 months they wanted to improve the use of data insights in business decision making, with the same number wanting to accelerate their organization's response to business and market changes. Around three in five (59%) respondents

said they most commonly use cloud marketplaces to procure third-party data, with more than half of that figure (33% overall) currently expanding plans. A further 21% of respondents were in the process of implementation. Cloud marketplaces were the most popular choice to purchase data among those polled, ahead of VARs (value-added resellers) and OEMs, managed service providers, SaaS consultants, or telcos.

For larger organizations, procuring new data or software comes with natural compliance risks. Outdated procurement processes can lead to shadow IT practices. "An unbending procurement process that doesn't adjust to new channels prompts

[23] "Cloud Migration Stats – 2022 Flexera State of the Cloud Report," Flexera, https://info.flexera.com/CM-REPORT-State-of-the-Cloud

[24] "Canalys Insights – Are cloud marketplaces worth the hype?", Canalys, https://www.canalys.com/insights/Cloud-marketplaces-as-a-channel-to-market

[25] Reduce Risk Exposure and Friction With Trusted Online Marketplaces, Forrester Consulting

even more employees to move outside of the traditional procurement processes, opening up organizations to increased risk," the Forrester report notes. Subsequently, the report revealed how online marketplaces promote procurement quality, speed, and safety. 86% of respondents said their organization was investing in new procurement channels to improve their business agility, while 78% agreed with the statement that online marketplaces for data and/or cloud software procurement decreases the organization's risk profile.

"Online marketplaces allow business teams to access a robust catalog of vendors and providers without sacrificing governance, risk, and compliance," the report concludes. "Firms that use online marketplaces today report an increased variety and options for providers, more trusted data, and an easier way to implement the right level of governance and control."

Enter AWS Data Exchange. AWS Data Exchange makes it easy to find, subscribe to, and use third-party data in the cloud. There is no other place where customers can find data files, data tables, and data application programming interfaces (APIs) from a vast portfolio of third-party data sets. Certain compliance risks can be mitigated through the use of Private Marketplaces (PMP)[26], which controls the products that users, from business users to engineering teams, can procure, and enables administrators to create and customize curated digital catalogs of approved products that confirm to an organization's policies.

Customers across many industries are seeing the benefits of utilizing AWS Data Exchange, augmenting their data with third-party data to further enhance the insights they can gain to help achieve their business outcomes. Here are a few examples:

Financial services

Goldman Sachs, being one of the world's leading global financial services providers, offers key services around investment banking, consumer and wealth management, and asset management. The company also has a global investment research division providing original, fundamental insights and analysis for clients in the equity, fixed income, currency and commodities markets. This analysis requires the use of third-party data, which needs to be ingested, parsed, and evaluated in order to deliver

[26] "Using third-party data from AWS Data Exchange", Amazon Web Services, https://catalog.us-east-1.prod.workshops.aws/workshops/e5548031-3004-49ad-89be-a13e8cd616f6/en-US/aws-data-exchange-governance/build-catalog-of-approved-products-via-private-marketplace

results to clients. This represents more than 400 external vendors supplying terabytes of data for this analysis. But this process was taking too long and using too many resources. Work was becoming duplicated. The company's leadership mandated a move to the cloud; in the words of managing director John Chappell[27], the company could not 'continue to effectively support clients just by continuing as is and adding more people to find, ingest, and process this growing list of data.'

"Data has powered the financial services industry for decades. Traditionally, quality of data, price, and usage rights were the key considerations for selection," added Chappell. "Now, how that data is delivered to us is as important, if not even more so." The move to the cloud necessitated the need for a different, cloud-based data catalog service – and this is where AWS Data Exchange comes in. To begin with, AWS Data Exchange is able to significantly lighten the workload across importing, processing and preparing the data, as well as simplifying access by enabling consumption directly in the cloud where it gets analyzed. Through AWS Data Exchange, Goldman Sachs can find and subscribe to a diverse range of data in one place, compliantly license third-party data at scale and manage entitlements across the organization.,

"AWS Data Exchange is a key component of Goldman Sachs's financial cloud strategy because it reduces friction for sourcing financial data from new and existing third-party providers and allows us to focus on delivering our core services and differentiated data analytics to better serve our clients," says Marco Argenti, Goldman Sachs co-CIO.

Media and entertainment

IMDb, the most popular and authoritative source for movie, TV and celebrity content, has a combined web and mobile audience of more than 200 million visitors per month. Yet there are many other services the company provides; in particular, services for media and entertainment organizations licensing a wide range of data to customers worldwide. This database consists of hundreds of millions of data points, from more than nine million TV and entertainment titles, to more than 11 million cast, crew and entertainment professionals' profiles, to metadata such as plot summaries, genres, credits, and release information.

[27] "Goldman Sachs and FactSet", Amazon Web Services, https://aws.amazon.com/partners/success/goldman-sachs-factset/

Use cases for this vast and authoritative database, which can be accessed in AWS Data Exchange[28], are myriad. For subscribers, organizations can leverage the rich metadata to group, analyze and extract value from content libraries, drive content purchases, subscriptions and retention, and power relevant, personalized content discovery features and recommendations. This can be through marrying first-party data – for instance, knowing that customer X loves a certain type of TV show – with third-party data of which shows rank highest among millions of viewers.

For a company such as Flixed, a provider of streaming availability data and service comparisons, their core proposition is to tell their customers where they can stream the movies and TV shows they want to watch. “Accurate and timely metadata about the titles in our catalog is critical,” explains Thenuka Karunaratne, founder and CEO at Flixed. “The combination of IMDb’s high quality data and AWS Data Exchange’s capabilities enable us to easily and quickly acquire that title metadata and use it in our data workflows running on AWS.”

Additionally, IMDb partners with AWS solution architects and account managers to help its customers find new ways of leveraging AWS services and technology.

Healthcare and life sciences

Vyaire Medical is a global healthcare company focused ‘exclusively on supporting breathing through every stage of life.’ This, practically, means the manufacture of products such as neonatal ventilators, airway management devices, and respiratory management items. Specifically, there are more than 27,000 unique products to more than 350,000 customers. During the Covid-19 pandemic, demand for ventilators understandably skyrocketed from hospitals all over the world. The company, through a series of mergers with other industry companies, had a complex technical ecosystem – and had reached the point where managing data was getting in the way of doing business.

The company realized that it needed more than its own first-party data to create better insights and decisions. Yet finding, licensing and ingesting this external data – search multiple sources to find the right data, sorting out the licensing and then getting the data ready for analysis – proved difficult and time-consuming for engineering teams. The solution to the problem was AWS Data Exchange[29]. “Going to

[28] "3 ways IMDB's third-party data enhances user engagement and experience", Amazon Web Services

[29] “Vyaire Uses AWS Data Exchange to Keep the World Breathing Better”, Amazon Web Services, https://aws.amazon.com/solutions/case-studies/vyaire-case-study/

the AWS Data Exchange catalog made me feel like a kid in a candy store," says Gopal Ramamurthi, vice president for analytics and global data management at Vyaire Medical. "There is a huge variety of data available to just grab and go; you can build, and experiment, and just start using what makes sense for your business."

Retail and CPG

Yum! Brands, headquartered in Kentucky, has 50,000 restaurants in more than 150 countries, and 2,000 franchisees. The company's brands include KFC, Taco Bell and Pizza Hut. Yum! Brands utilizes a wide range of internal data, from historical sales to attributes about their stores. This ranges from the type of store – drive-thru, dine-in, mall – to seating capacity, to how operations are run in that store. In terms of expansion, this means selecting physical locations for new restaurants. To do this, Yum! Brands needed third-party data that shows which geographical locations are of interest and why. The answer was through utilizing Foursquare's Point-of-Interest data through AWS Data Exchange. In a webinar[30], Nikhil Jain, senior director of decision sciences at Yum! Brands, explained that "every brand in the company has their own strategy… for us to know what businesses are where is key for the success of our models." The Point-of-Interest data is utilized along with sociodemographic data and traffic data. This, practically, relates to 250 meter x 250 meter grids. "The precision of the data has to be really good at such a low level," added Jain. "If [not], then operations are not going to be good. That's something we depend on Foursquare and AWS to provide."

The benefits, therefore, are improved customer experience with convenient restaurant locations, and better decision-making based on data and insights, versus just human experience or instinct. "Working with Foursquare and AWS Data Exchange has enabled us to seamlessly procure the exact Point-of-Interest data we needed across dozens of markets, and all delivered to us directly in the cloud alongside our data analytics tools," added Jain[31].

These examples show how brands who are leaders in their fields are able to answer key business challenges through supplementing their first-party data with third-party data – and how AWS Data Exchange has been able to help them. In the next chapter,

[30] "How Yum! Brands uses location data from Foursquare to make smarter decisions", Amazon Web Services, https://pages.awscloud.com/adx-h2-3rd-party-data-yum-brands-ty.html

[31] "CPG and Retail Data Sets on AWS Data Exchange," Amazon Web Services, https://aws.amazon.com/data-exchange/cpg-and-retail/

we will look at why AWS Data Exchange is able to help your organization and the attributes which make it stand out.

Chapter 4: AWS Data Exchange – a service for organizations looking to buy and sell third-party data

The range and volume of data available today – both first-party generated data and third-party data – enables companies to improve customer experiences and innovate faster than ever using insights gained through analytics and machine learning. AWS Data Exchange streamlines all third-party data consumption, from existing subscriptions to future purchases, speeding up time-to-insights and future proofing the growing analytics functionality at the company. AWS Data Exchange makes the process of getting third-party data directly into the applications where it's needed faster and easier, allowing teams across the company – from data scientists, to IT, to finance – to focus on more strategic work.

AWS Data Exchange aims to make the world's third-party data easy to find in one data catalog, simple to subscribe to with consistent pricing options, and seamless to use with other AWS services. Before we explain how to get started with AWS Data Exchange, let's look at a few examples of integrations with AWS services, as well as features which promote ease of use:

Identity and access management (IAM)

To perform any operation in AWS Data Exchange, AWS Identity and Access Management requires that you authenticate that you're an approved AWS user. After your identity is authenticated, IAM controls your access to AWS with a defined set of permissions on a set of operations and resources. Account administrators can use IAM to control the access of other users to the resources that are associated with their account.

AWS Data Exchange can simplify permissions through managed policies, a standalone policy that is created and administered by AWS. One particularly useful category of AWS managed policies are those designed for job functions. The intent is to make granting permissions for these common job functions easy. Teams can get started quickly with AWS managed policies, which cover common use cases and are available in your AWS account.

Data ingestion

Data ingestion is often another complex and time-consuming process for teams. AWS offers auto-export as a way to make data ingestion faster and easier. If a subscriber receives updated data on AWS Data Exchange, which is then exported into Amazon S3 buckets, they can then 'set and forget' their export preferences, and AWS Data Exchange automates the delivery of any new revisions to the S3 buckets specified. For users who consume third-party data directly into their S3 bucket(s), AWS Data Exchange is able to consolidate ingestion across data providers so users can receive their data using a single API.

APIs

APIs are the building blocks which allow the integration of application software – in other words, an accessible way to extract and share data within and across organizations. With AWS Data Exchange, subscribers don't have to build bespoke data pipelines for every use case to retrieve small amounts of data from various sources. They can simply use their AWS IAM credentials and AWS SDKs to call data APIs from dozens of data providers.

How To Get Started With AWS Data Exchange

Any AWS account holder can use AWS Data Exchange, but they need to first create an AWS IAM user before subscribing.As a subscriber, you can find and subscribe to thousands of products from qualified data providers; then, you can use the AWS Data Exchange console or API to view, manage and access data sets for use across a variety of AWS services across analytics and machine learning.

AWS Data Exchange dovetails with various other AWS services. These include, but are not limited to:

Amazon S3**: Amazon Simple Storage Service (S3) is an object storage service offering industry-leading scalability, data availability, security and performance.** AWS Data Exchange allows providers to import data files from their Amazon S3 buckets or local file systems, with subscribers being able to export these files to Amazon S3. Subscribers can also directly access and use providers' Amazon S3 buckets. Amazon Redshift: Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the AWS Cloud. An Amazon Redshift data warehouse is a collection of computing resources called nodes, which are organized into a group called a cluster. Each cluster runs an Amazon Redshift engine and contains one or more databases.

AWS Lake Formation**:** AWS Lake Formation easily creates secure data lakes, making data available for wide-ranging analytics. Users can create, administer, and protect data lakes using familiar database-like features quickly, simplify security management and governance at scale, and break down data silos and make all data discoverable with a centralized data catalog.

Once you have signed up for AWS and created an AWS IAM user, you then take the following steps to subscribe to data products on AWS Data Exchange:

You can find products, which are published both on AWS Data Exchange and AWS Marketplace, and review the associated public or custom offers and product details. If you subscribe to a paid product, you are billed on your AWS bill, and you get access to the entitled data set. Data providers can provide a product to subscribers that are either not available to the general public, or at terms that are different from the publicly available offer terms. This is known as a private offer and can include price, duration, or payment schedule among others. If you are unsure as to which data product would fit best, then as a subscriber you can request personalized recommendations from the AWS Data Exchange Data Discovery Team.

As above, the data set is by no means siloed, meaning the subscriber can interact with it across their AWS environment or migrate it elsewhere. You can take any of the following actions: export the associated files to your Amazon S3 bucket, or locally through a signed URL; call the Amazon API Gateway API; query the Amazon Redshift data share; or access or query the provider's Amazon S3 data access or AWS Lake Formation data lake. Subscribers who have existing data subscriptions can migrate them to AWS Data Exchange through Bring Your Own Subscription (BYOS) functionality.

This chapter has outlined the key benefits of AWS Data Exchange, how it can integrate with other AWS services, and how to subscribe to data products on AWS Data Exchange. The final chapter outlines several workshops which will help users get hands-on with AWS Data Exchange, and how data from the product can be visualized and analyzed through AWS tools and services.

Chapter 5: Workshops

To find out more about how to most effectively utilize AWS Data Exchange, AWS has put together a series of optional workshop modules which contain several self-service labs to help users understand the relationship between third-party data and AWS services. The majority of labs can be completed in under 45 minutes.

Business intelligence users are recommended to explore the Data Visualization and Analytics labs in particular, while data analysts or data scientists are recommended Analytics and machine learning labs.

The AWS Data Exchange workshops are as follows:

- AWS Data Exchange overview: This module helps users understand important concepts about AWS Data Exchange, and learn how to browse the AWS Data Exchange catalog
- Data subscription and export workflow: Users will initially learn how to subscribe to a data product. Then with AWS Data Exchange for Data Files, users will learn how to perform a one-time export of the complete dataset; set up automatic export for all future revisions; set up notifications with Amazon EventBridge; and as an optional module, set up file transfer workflow via SFTP (Secure File Transfer Protocol). With AWS Data Exchange for Amazon Redshift, users can learn how to directly query data from datashare products, and directly query data from datashare products using Amazon Redshift Serverless. With AWS Data Exchange for APIs, users can learn how to make API calls to the data product
- Visualizing data: Users will learn how to visualize data via tools such as Amazon Quicksight and Jupyter notebooks
- Performing analytics on your data: Users will learn various methods of how to perform analytics on top of the data. The labs involve doing visual data preparation with AWS Glue DataBrew; populating metadata in AWS Glue Data Catalog; running ad hoc queries on top of the data using Amazon Athena; and joining and transforming data with Amazon Redshift Spectrum
- Machine learning: Data from AWS Data Exchange can be used with various AWS services to train ML models. Users will learn how to train a machine learning model, build a recommendations engine with Amazon Personalize, and perform inference on an ML model

- AWS Data Exchange governance: Users will learn in these labs how to set up guardrails around their AWS Data Exchange experience. This includes setting up a Private Marketplace to enable curated catalogs and approval workflows, and using License Manager to manage entitlements to data with other accounts in your organization

- Publish and manage data products: This module and workflow is aimed at data providers. In these labs users will learn how the end-to-end processes of becoming a data provider, creating data sets that go into products, and listing products and extending private offers to specific customers

- AWS Marketplace Discovery API and service integrations: The AWS Marketplace Discovery API allows AWS partners to implement a frictionless discovery experience of AWS Data Exchange listings on their web properties. This enables direct their customers to the most up-to-date and relevant AWS Marketplace pages to purchase the products they need. Users will learn an overview of Discovery API, how to explore the Discovery API with Python, and create a Discovery API-powered web application.