Open Data on AWS

March 30, 2022

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.

Contents

Abstract

Open Data is data that can be freely used, re-used and redistributed by anyone—subject only to the requirement to attribute data and share it equitably. Open Data is structured, machine-readable, open-licensed, and well-maintained.

According to a 2014 McKinsey report on government Open Data, governments around the world are providing access to data and publishing that data on their websites for the purpose of research, analytics, and application development.

The McKinsey report further suggests that governments can promote Open Data and unleash more than $3 trillion in economic value.

There are many reasons why Open Data initiatives have not yet scaled effectively. In a recent report, Opendatabarometer.org suggests that governments have published only 7% of available data as Open Data—and typically only on their own websites.

Making Open Data available in a cloud environment—and especially a registry of Open Data on Amazon Web Services (AWS)—helps scale Open Data initiatives.

This paper examines the current state of Open Data initiatives. It discusses the various challenges that Open Data projects can encounter, and how cloud strategy can help meet these challenges. The primary focus of this paper, however, is to provide a perspective on, and guidance around, the legal framework in India that addresses Open Data policy.

This is important, as there are a unique set of compliance requirements for use of Open Data in India, including the *National Data Sharing and Accessibility Policy (NDSAP)* and the *Right to information Act, 2005*, that must be followed to implement Open Data solutions in India.

This paper provides practical guidelines covering the full lifecycle of Open Data management—from selecting the dataset and pre-processing the data to publishing and post-processing for legal compliance.

It describes the current infrastructure of Open Data frameworks on AWS and presents a recommended architecture that can help customers develop a secure Open Data pipeline that aligns with their compliance requirements for handling personally identifiable information (PII).

Introduction

Open Data is data that can be freely used, re-used and redistributed by anyone—subject only to the requirement to attribute data and share it equitably. Open Data is structured, machine-readable, open-licensed, and well-maintained.

Key characteristics of Open Data

Key characteristics of Open Data are as follows:

- **Complete** – Agencies should release all public data that are not subject to privilege, security, or privacy limitations.
- **Primary** – Data should not be aggregated or modified; it should come from its primary source to ensure integrity.
- **Timely** – Data should be made available as soon as possible to ensure the fairness of Open Data opportunity.
- **Machine-readable** – Data should be in machine-readable format to foster automation.
- **Available** – Data should be available to anyone.
- **Non-exclusive** – All parts of an Open Data dataset should be unconditionally available. No exclusive access to any part of the data.
- **Open Use and distribution** – Data should not be subject to any copyright, trademark, trade secret, or patent protections.
- **Security** – Data is subject to the three aspects of information security: confidentiality, integrity, and availability. Raw data needs to be protected from unauthorized manipulation.
- **Free** – Data should be available at either no cost, or as low a cost as is reasonably possible.
- **Data Ownership** – Data should have ownership properties that are maintained by clear policy, and those properties must define access, co-ownership, and transformation rights.

Advantages of Open Data

Governments around the world have recognized Open Data as one approach to increase transparency, advance citizen engagement and economic welfare, as well as improve policy making and public decision making. The initiatives of OGD (Open Government Data) are increasing steadily, and this advancement has introduced both opportunities and challenges.

With Open Data, governments can provide greater transparency to the work of all public agencies, potentially fostering greater interest and trust in public affairs. Public sector agencies can also use Open Data to store their information in industry-standard data formats, thus reducing the cost of processing that data. Developers can use this structured data to develop applications more quickly.

The three stages of Open Data development

There are three major stages in the development and distribution of Open Data: creation, exploration, and reuse.

At the creation stage, Open Data involves integrating data from different datasets to produce it in an industry-standard data format. At the exploration stage, data is represented visually to allow it to be understood and used more effectively. The third stage—re-use—requires modifying data to allow it to be re-used more effectively.

The success of Open Data relies on taking an integrated approach to meeting a number of challenges, such as:

- **Data selection** – The organization needs to identify the data that can be published without compromising public security and privacy. Check out *Appendix 1 – Open Data types for consideration* for examples of typical data that governments may be able to consider making open.
- **Data heterogeneity** – Data may be stored in structured or unstructured formats, which may or may not be machine readable. The data needs to be standardized before being published as Open Data.
- **Non-uniform data access** – There can be a lack of proper standards dealing with availability of datasets across government services.
- **Data security** – Proper data security measures should be in place.

- **Data quality** – Parameters of quality of the dataset should be standardized across data sources. These parameters can indicate areas of improvement and help safeguard the quality of the dataset published on the portal.
- **Data pre-processing** – Data should be pre-processed to ensure that the data meets applicable compliance requirements before it is published as Open Data.
- **Data ownership** – The publisher of the data owns the dataset. A data governance structure that defines who is responsible for publishing, maintaining, and updating should be put in place.

Data protection legal framework in India

Data protection refers to the set of privacy laws, policies, and procedures that aim to minimize intrusion into one's privacy caused by the collection, storage, and dissemination of personal data. Personal data generally refers to the information or data which relate to a person who can be identified from that information or data—whether the data is collected by any government, private organization, or agency.

What data can be made publicly available will be limited by legal requirements, which differ across countries and jurisdictions As an example, we examine how Open Data policy in India is governed by the following legal instruments:

- National Data Sharing and Accessibility Policy (NDSAP)
- Right to information Act, 2005
- Information Technology Act, 2002
- Information Technology (Reasonable security practices and procedures and sensitive personal data or information) Rules, 2011

India's Information Technology Rules, 2011 makes agencies responsible in cases where the following personal sensitive data (also known as personally identifiable information (PII)) is inadvertently disclosed.

- Passwords
- Financial information, such as bank account, credit card, debit card, or other payment instrument details
- Physical and mental health condition
- Sexual orientation

- Medical records and history
- Biometric information

Traditional approach of Open Data

Generally speaking, Open Data is information that has been published on government-sanctioned portals. In the best case, this data is structured, machine readable, license-free, and well maintained.

According to opendatabarometer.org, only a small percentage of government data is currently published as Open Data. It reports that "only 7% of the data is fully open, only one of every two datasets is machine readable, and only one in four datasets has an open licence."

In addition, the report also suggests that Open Data is often fragmented and published on multiple government portals—and public awareness of the availability of the data is low.

As a result, to make effective use of the data, researchers and analysts have to download Open Data and then process it. A great deal of effort goes into accessing the data, cleaning it up, and making it ready for analysis. This creates a barrier to the usability of Open Data, as it is published today.

Open Data on the Cloud

The traditional approach to working with data is to download it first to a data centre, networked server, desktop, or laptop computer (depending on the size of the data).

Consider an example where the requirement is to work with a 10 TB dataset. In a traditional data centre, the first step would be to figure out how to get the 10 TB of data to a storage location where you had the computing power (and storage capacity) to work with it.

In a cloud environment, the challenge is flipped. Instead of moving data around where your algorithm is, you can move the algorithm to where data is by launching an instance within a virtual private cloud (VPC).

You can scale up as your compute needs change. If you need to do large-scale analytics, you can launch as many instances as needed and shut them down when your

analysis is done. You can scale up and down much more efficiently than if you were moving data around.

Advantages of Open Data on AWS

In this section, we discuss the advantages of using Open Data on AWS.

Global community of users

When you share data on AWS, you make it available to a large and growing global community of developers, start-ups, and enterprises. Members of the community write and share code with one another, also sharing best practices on how to use data effectively.

New services and tools

Data shared on AWS becomes more useful as new features and services are released. AWS released tools such as AWS Glue DataBrew and Amazon SageMaker Data Wrangler to help work with data more efficiently.

Reduce time to insight

AWS data analysis tools are designed to allow customers to focus more quickly on analysis, without having to do a lot of pre-work downloading or cleaning data.

Customers can store structured and unstructured data (images, video, satellite imagery, call transcripts, and so on) in Amazon Simple Storage Service (Amazon S3). If it is stored in machine-readable format, a user can source relevant data automatically and then transform it to their purpose and build use cases.

For example, the European Satellite Agency (ESA) has published Sentinel data in cloud-optimized GeoTIFF (COG) format, which can be directly read programmatically for further processing. A cloud-optimized GeoTIFF (COG) is a regular GeoTIFF file, aimed at being hosted on a web server, with an internal organization that enables more efficient workflows. The file does this by leveraging the ability of clients that are issuing HTTP GET range requests to ask for just the parts of a file they need. It is a widely used data or dataset for agriculture and urban analytics use cases.

Lower cost of research

Researchers can analyze data shared on AWS without paying for their own stored copy. They only pay for the compute they use, and do not need to purchase storage to start a

project. That lowers the cost of research. Researchers can also perform research faster, because they do not have to spend time on acquiring data.

Registry of Open Data on AWS

The Registry of Open Data on AWS is a website that is designed to help researchers find datasets that are publicly available through AWS. When data is shared on AWS, anyone can analyze it without needing to download or store it, which allows users to spend more time on analysis, rather than acquisition. The Registry of Open Data on AWS (RODA) is available at https://registry.opendata.aws/

RODA allows you to search for datasets by keyword and tags for common types of data, such as genomic, satellite imagery, and transportation. If you have data that you would like to put into the registry, you can do so by adding it through GitHub. Step-by-step instructions on how to add data are as follows.

How to add data to the Registry of Open Data

Datasets in this registry are available through AWS, but they are not provided by AWS; these datasets are owned and maintained by a variety of government organizations, researchers, businesses, and individuals.

The first step is to set up the AWS resource for delivering the data. Available AWS resources include any of the following options:

- Amazon CloudFront distribution
- Amazon Relational Database Service (Amazon RDS) database snapshot
- Amazon Elastic Block Store (Amazon EBS) snapshot
- Amazon S3 bucket
- Amazon Simple Notification Service (Amazon SNS) topic

Data has to be pre-processed, so that it is in machine-readable data format, which may or may not be limited to CSV, GeoTIFF, shapefile, and other standard formats.

The second step is to create metadata for each dataset, saved in a YAML file. The YAML file is saved in the GitHub repository. Saved YAML files periodically get crawled, and an entry is created at Registry of Open Data on AWS. This provides three benefits:

1. Inclusion in the Registry of Open Data on AWS

2. A hosted YAML file listing of all the datasets
3. Hosted YAML files for each dataset

Detailed instructions for making a YAML entry is provided at the Registry of Open Data on AWS GitHub repository.

Access to datasets on the registry is provided through Amazon Resource Names (ARNs). ARNs uniquely identify AWS resources. An ARN is required when you need to specify resources unambiguously across all of AWS, such as in an API call.

Optionally, you can apply for the AWS Open Data Sponsorship Program. The AWS Open Data Sponsorship Program covers the cost of storage and data transfer for a period of two years for publicly available high-value cloud-optimized datasets. It is available to data providers who seek to do the following:

- Democratize access to data by making it available for analysis on AWS
- Develop new cloud-native techniques, formats, and tools that lower the cost of working with data
- Encourage the development of communities that benefit from access to shared datasets

Program enrollment is renewed after two years. For more details on applying, see the Open Data Sponsorship Plan.

Data governance with Amazon Macie

How do you validate and prove that there is no PII in your data storage, such as in an S3 bucket? If there is PII, how do you know what information it is, as well as where it is? This is exactly where Amazon Macie can help the Open Data publisher.

Amazon Macie is a cloud security service that uses machine learning to identify and protect sensitive data that is stored in the AWS public cloud.

Amazon Macie automatically and continually discovers sensitive data, such as PII or intellectual property, in the stored open data in Amazon S3. You can configure a one-time job or a scheduled job to scan for the sensitive information in Open Data.

AWS infrastructure for creating an open data pipeline

AWS provides multiple services that customers can use in tandem to create an open data pipeline and which are designed to ensure that the data that is made public does

not contain any personally sensitive data. A proposed reference architecture is shown in the following diagram.

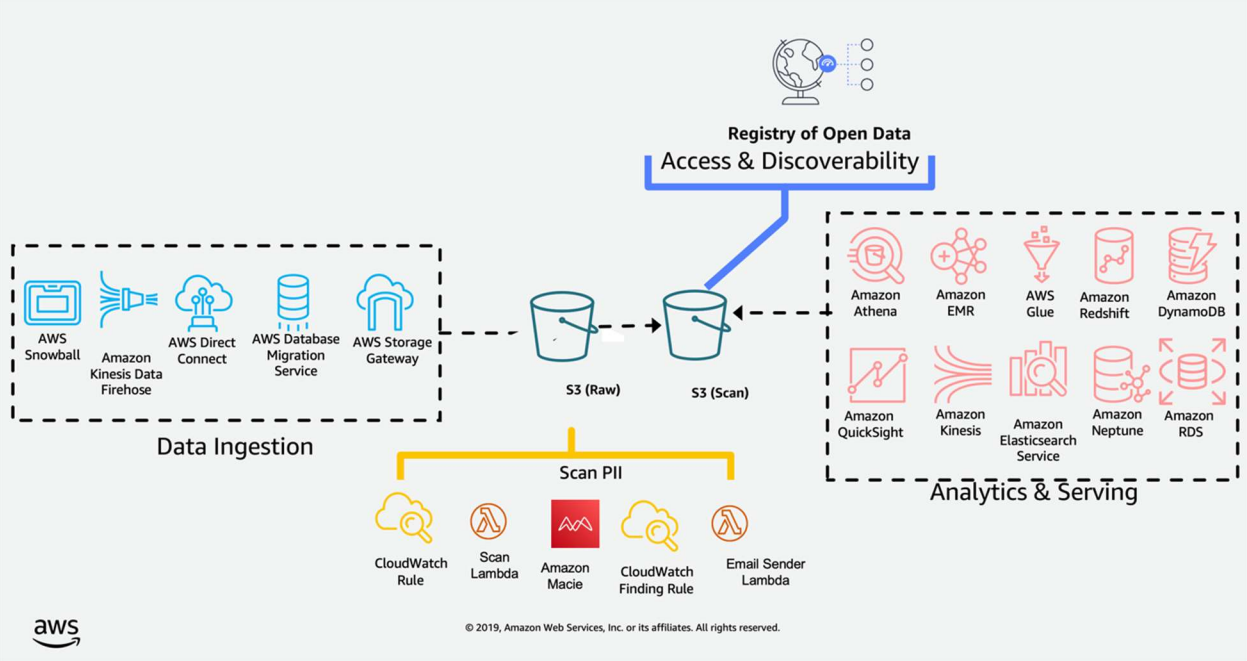

Figure 1: Open data pipeline reference architecture

Data ingestion

AWS provides multiple service options to get data on AWS, either from on-premises or online. Through the AWS Storage Gateway, on-premises data storage can be augmented with Amazon S3 storage. An Amazon Kinesis Data Firehose can help integrate with streaming data to Amazon S3. In our proposed architecture, we create a bucket in Amazon S3 to store the ingested raw data. This bucket will store a copy of the dataset we want to make open in machine-readable format.

Scanning

We would like to make sure the data we have made open is PII compliant. To accomplish that, we propose a pre-processing step of scanning data by using the Amazon Macie service. Macie sends a notification if PII is found, so that customers review the data against their compliance requirements. If no PII is found, scan data is saved to another S3 bucket. This scan dataset is registered with the Registry of Open Data on AWS (RODA) and made available for others to access. This step can help customers to ensure that their data conforms with applicable PII compliance

requirements, like a requirement that name, Social Security number, passport number, bank details, and credit card number are not made publicly available.

Analytics and serving

Researchers and application developers should be able to analyze available Open Data with the click of a button. The AWS Analytics & Serving stack provides AWS services that can directly integrate with Open Data resources. Amazon EMR, Amazon Athena, Amazon DynamoDB, and Amazon Redshift can be used directly to analyze the data that has been made open.

Access and discovery

Access and discovery of Open Data is enabled by the Registry of Open Data on AWS. Legally compliant data in S3 buckets is registered on the registry, which can be searched and discovered by researchers and analysts.

AWS Data Exchange compared to Registry of Open Data on AWS

AWS Data Exchange is a service that makes it easy for AWS customers to securely find, subscribe to, and use third-party data in the cloud. After you are subscribed to a data product, you can use the AWS Data Exchange API to load data directly into Amazon S3 and then analyze it with a wide variety of AWS analytics and machine learning services.

For data providers, AWS Data Exchange makes it easy to reach AWS customers who are migrating to the cloud by removing the need to build and maintain infrastructure for data storage, delivery, billing, and entitling.

There are five key differences between AWS Data Exchange and the Registry of Open Data on AWS:

- AWS Data Exchange supports both free and commercial data products, with any applicable commercial fees applied to a customer's AWS invoice, whereas the Registry of Open Data on AWS gives customers access to a curated list of free and open datasets.
- AWS Data Exchange requires customers to explicitly agree to the Data Subscription Agreement outlining the terms that the data provider set when publishing their product, whereas the data on the Registry of Open Data on AWS does not have terms of use.

- Customers must use the AWS Data Exchange API to copy data from AWS Data Exchange to their desired S3 location, whereas the datasets on the Registry of Open Data on AWS can be accessed by S3 APIs.
- AWS Data Exchange delivers data providers access to daily, weekly, and monthly reports that detail subscription activity, whereas with Registry of Open Data on AWS, data providers must analyze their own logs to track usage of data.
- To become a data provider on AWS Data Exchange, qualified customers must register as a data provider on AWS Marketplace to be eligible to list both free and commercial products, whereas any customer can add free data to the Registry of Open Data on AWS via GitHub and may apply to the AWS Public Dataset Program for AWS to sponsor the costs of storage and bandwidth for select open datasets.

Conclusion

Open Data projects need collective and continuous efforts, and cloud-friendly Open Data policy is the key to providing that foundation. There are issues on both the supply side and consumption side of data availability.

The discussion in this whitepaper assesses the challenges of Open Data initiatives, followed by recommendations for a range of best practices and cloud infrastructure to help scale. These principles may constitute robust Open Data initiatives. These recommendations include policy making, technical implementation, data quality, continuous improvement, and legal framework.

Open Data works better on a larger scale. It also produces better results when it provides a consolidated access and discovery mechanism. Governments around the world are looking for success stories that foster Open Data initiatives, and by understanding cloud infrastructure and its strong community of users and easily available compute resources, we may resolve the outstanding challenges of Open Data projects by successfully exploiting its evident benefits.

Appendix 1 – Open Data types for consideration

Serial no.	Public content domain	Examples
1	Geographic information	Cartographic information
		Land use and land cover
		Spatial data/geographic coordinates
		Administrative and political boundaries
		Elevation data
2	Meteorological and environmental information	Oceanographic data
		Hydrographic data
		Environmental (quality) data
		Atmospheric data
		Meteorological (weather) data
3	Economic and business information	Financial information
		Company information
		Economic and statistics
		Industry and trade information
4	Social information	Demographic information
		Attitude surveys
		Data on health and illness
		Education and labour statistics
5	Traffic and transport information	Transport network information
		Traffic information
		Transport statistics

Serial no.	Public content domain	Examples
		Car registration data
6	Tourism and leisure information	Hotel information
		Tourism statistics
		Entertainment (local and national)
7	Agriculture, farming, forestry, and fisheries information	Cropping/land use data
		Farm incomes/use of resources
		Fish farming/harvest information
		Livestock data
8	Natural resource information	Biologic and ecologic information
		Energy resource and consumption information
		Geological and geophysical information
9	Legal system information	Crime and conviction data
		Laws
		Information on rights and duties
		Information on legislation
		Information on judicial decisions
		Patent and trademark information
10	Scientific information and research data	University research
		Publicly funded research institutes
		Governmental research
11	Educational content	Academic papers and studies
		Lecture material

Serial no.	Public content domain	Examples
12	Political content	Government press releases
		Local and national proceedings of governments
		Green papers
13	Cultural content	Museum material
		Gallery material
		Archaeological sites
		Library resources
		Public service broadcast archives
		Other public archive

Source: Adapted from Pira, PSINet, and other studies

Appendix 2 – Legal framework for data protection

National Data Sharing and Accessibility Policy (NDSAP) 2012

The National Data Sharing and Accessibility Policy (NDSAP) is applicable to all shareable, non-sensitive data available either in digital or analog forms, but generated using public funds by various ministries, departments, subordinate offices, organizations, and agencies of the government of India, as well as of the states.

The objective of this policy is to facilitate access to Government of India–owned shareable data through a wide area network, thereby permitting a wider accessibility and usage by public. The principles on which data sharing and accessibility need to be based include: openness, flexibility, transparency, quality, security, and efficiency.

Right to Information Act, 2005

The Right to information Act is an act of the Parliament of India, which sets out the rules and procedures regarding citizens' right to information.

Information Technology Act, 2002

The Information Technology Act, 2002 stipulates various provisions in order to provide for the legal framework, so that legal sanctity is accorded to all electronic records and other activities carried out by electronics means. The Act legalizes the use of digital signatures and authentication of electronic records.

Information Technology (Reasonable security practices and procedures and sensitive personal data or information) Rules, 2011

The Rules only deal with protection of "Sensitive personal data or information of a person," which includes personal information such as the following:

- Passwords
- Financial information, such as bank account or credit card or debit card or other payment instrument details
- Physical, physiological, and mental health condition
- Sexual orientation
- Medical records and history
- Biometric information

The Rules provide the reasonable security practices and procedures, which the corporate body or any person who on behalf of the body corporate collects, receives, possesses, stores, deals, or handles information is required to follow while dealing with "Personal sensitive data or information." In case of any inadvertent disclosure, the corporate body or any other person acting on behalf of the corporate body may be held liable to pay damages to the person so affected.

Contributors

Contributors to this document include:

- Sanjiv Kumar Jha, Principal Smart Infra SA, Amazon Web Services

Document revisions

Date	Description
July 2021	First Publication.