Data masking for data in AWS RDS [closed]

Asked 7 years, 1 month ago Modified 6 years, 9 months ago

Viewed 10k times

9

Closed. This question does not meet Stack Overflow guidelines. It is not currently accepting answers.

Closed 6 years ago.

- This question does not appear to be about a specific programming problem, a software algorithm, or software tools primarily used by programmers. If you believe the question would be on-topic on another Stack Exchange site, you can leave a comment to explain where the question may be able to be answered.
- We don't allow questions seeking recommendations for software libraries, tutorials, tools, books, or other off-site resources. You can edit the question so it can be answered with facts and citations.

Improve this question

I have an AWS RDS (AuroraDB) and I want to mask the data on the DB. Does Amazon provides any service for data masking?

I have seen RDS encryption but I am looking for data masking because the database contains sensitive data. So I want to know is there any service they provide for data masking or is there any other tool which can be used to mask the data and add it manually into the DB?

A list of tools which can be used for data masking is most appreciated if any for mine case. Because I need to mask those data for testing as the original DB contains sensitive information like PII(Personal Identifiable information). I also have to transfer these data to my co-workers, so I consider data masking an important factor.

Thanks.

data-masking

Share

Improve this question

Follow

edited Nov 8, 2017 at 1:44

asked Nov 7, 2017 at 12:26

Neron Joseph

2,285 ● 5 ● 30 ● 52

You can publish the data to an SNS topic, and have the topic mask the sensitive data for you, in real time. This data protection feature was first introduced to SNS in September 2022 in public preview, then announced as generally available (GA) in November 2022. This feature supports sensitive data auditing, blocking, masking, and redaction. For

more information, see: docs.aws.amazon.com/sns/latest/dg/message-data-protection.html – Otavio Ferreira Mar 4, 2023 at 1:53

2 Answers

Sorted by: Highest score (default)

2

This is a fantastic question and I think your pro-active approach to securing the most valuable asset of your business is something that a lot of people should heed, especially if you're sharing the data with your co-workers. Letting people see only what they need to see is an undeniably good way to reduce your attack surfaces. Standard cyber security methods are no longer enough imo, demonstrated by numerous attacks/people losing laptops/usbs with sensitive data on. We are just humans after all. With the GDPR coming in to force in May next year, any company with customers in the EU will have to demonstrate privacy by design and anonymisation techniques such as masking have been cited as way to show this.

NOTE: I have a vested interest in this answer because I am working on such a service you're talking about.

We've found that depending on your exact use case, size of data set and contents will depend on your masking method. If your data set has minimal fields and you know where the PII is, you can run standard queries to replace sensitive values. i.e. John -> XXXX. If you want to maintain some human readability there are libraries such as Python's Faker that generate random locale based PII

you can replace your sensitive values with. (PHP Faker, Perl Faker and Ruby Faker also exist).

DISCLAIMER: Straight forward masking doesn't guarantee total privacy. Think someone identifying individuals from a masked Netflix data set by cross referencing with time stamped IMDB data or Guardian reporters identifying a Judges porn preferences from masked ISP data.

Masking does get tedious as your data set increases in fields/tables and you perhaps want to set up different levels of access for different co-workers. i.e. data science get lightly anonymised data, marketing get a access to heavily anonymised data. PII in free text fields is annoying and generally understanding what data is available in the world that attackers could use to cross reference is a big task.

The service i'm working on aims to alleviate all of these issues by automating the process with NLP techniques and a good understanding of anonymisation maths. We're bundling this up in to a web-service and we're keen to launch on the AWS marketplace. So I would love to hear more about your use-case and if you want early access we're in private beta at the moment so let me know.

edited Nov 8, 2017 at 9:32

answered Nov 8, 2017 at 9:20

rimeice
141 • 5

0

If you are exporting or importing data using CSV or JSON files (i.e. to share with your co-workers) then you could use FileMasker. It can be run as an AWS Lamdbda function reading/writing CSV/JSON files on S3. It's still in development but if you would like to try a beta now then contact me. Disclaimer: I work for DataVeil, the developer of FileMasker.

Share Improve this answer

Follow

answered Nov 9, 2017 at 5:18

uuqpo
16 • 1