

Using Data for Community Growth

Arthur Diniz



whoami

- Debian Contributor since 2019
- Debian Maintainer
- Systems Development Eng @ AWS
- arthurbdiniz
- Born in Brazil 
- Living in Ireland 
- Kite Surfing



Slides



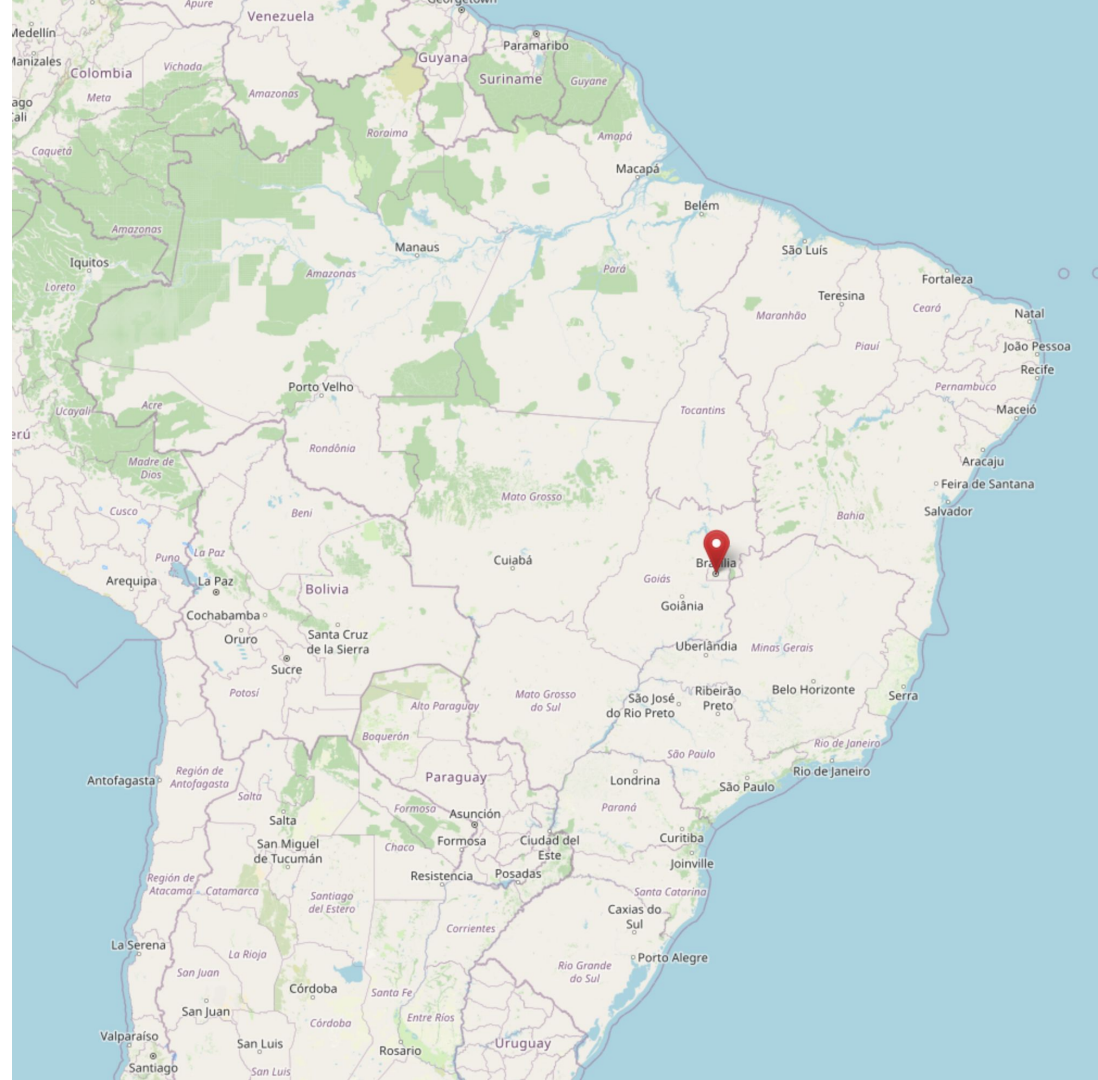
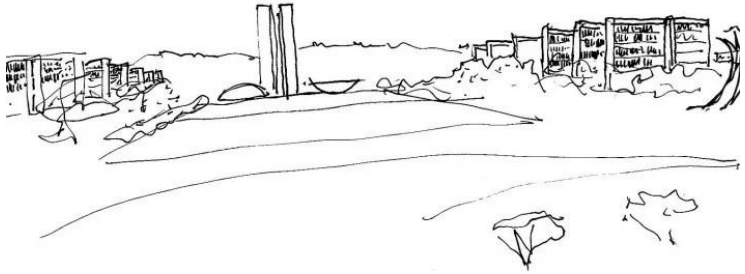
<https://arthurbdiniz.com/slides/using-data-for-community-growth.pdf>

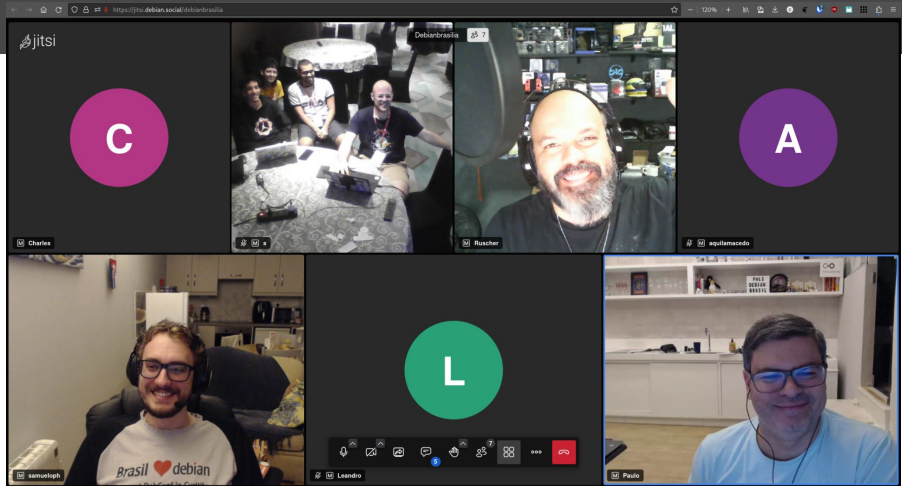
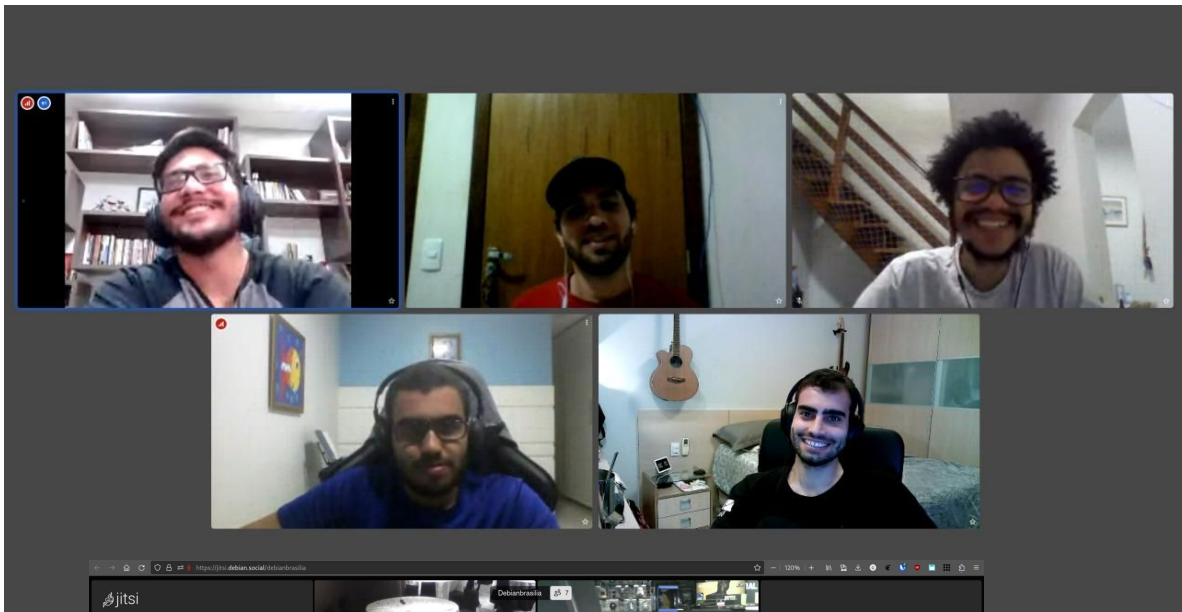
Agenda

- Debian Brasilia Local Community
- Tools
- ETL
- Data Lake and Data Warehouse
- Visualization
- Next Steps



debianbrasil





Pad

O QUE É ESSE PAD?

Criamos o pad para organizar as contribuições e salvar anotações interessantes. Na seção WORKING estão listados os pacotes que estão sendo trabalhados, na seção RFS estão os pacotes finalizados e que estão prontos para serem revisados (quando colocar é legal avisar lá no grupo para as pessoas saberem que o pad foi atualizado). Por fim, nas ANOTAÇÕES estão coisas aleatórias, ideias que foram discutidas nos encontros e alguns links.

<https://lazyfrosch.github.io/training-debian-packaging/dev/gbp/>

kubernetes wiki: <https://salsa.debian.org/debian-brasilia-team/docs/-/wikis/Kubernetes-Packages>

-

RFS (Request For Sponsorship)

- [nome] [Merge Request / algum link para avaliação]
- |-> [Ricardo] <https://salsa.debian.org/progronauta/python-colored>
- |-> [Ricardo] <https://salsa.debian.org/progronauta/texttable>

ANOTAÇÕES

- <https://bugs.debian.org/release-critical/debian>
- <https://dep-team.pages.debian.net/deps/dep3/>
- `dpkg -c # mostrar conteúdo do pacote`
- `dpkg-deb --extract ../[pacote].deb tmp/ # extrair pacote`
-
- `lintian -EI --pedantic ../<pacote>.dsc`
- `lintian -EI --pedantic`
- `egrep -sriA25 '(public dom|copyright)' | less` (comando do Eriberto para verificar as licenças/copyright)
- gbp:
- `gbp dch --team # gerar o debian changelog a partir dos commits (--team cria uma linha mostrando que é Team Upload)`
- `gbp buildpackage --git-builder=sbuild --git-dist=sid` (buildar utilizando o sbuild)
- `gbp pq import --time-machine=3`
- `gbp pq rebase`
- `gbp pq import ->` criar a branch
- `gbp pq export`
- `Autopkgtest:`
- `autopkgtest -BU -s <pacote>.changes -- schroot chroot:[nome do chroot]`
- Schroot:
- `schroot --begin-session --chroot chroot:[nome do chroot] --session-name [nome da sessão]`

board.debianbsb.org

Debian Brasilia / docs / Issue Boards

Debian Brasilia

Search



Show labels



Edit board

Create list



Open

TODO 41

Package pymongo

GCES

#114

docker-compose

#14

Package paramspider

NEW issues-preventing-migration

#185

Package exiflooter

NEW issues-preventing-migration

#144

Package waymore

NEW need-source-upload

#176

Package raven

NEW need-source-upload

#143

Package pyrandom2

GCES

#106

python-braintree

Doing 16

Package python-static3

#75

Package ruby-spy

GCES

#87

Package arpswitch

NEW

#97

Package pyparted

GCES

#104

Package ruby-cabypara

GCES

#117

Package ruby-factory-bot-rails

GCES

#119

Package jamulus

GCES

#122

Package ruby-kramdown-rfc2629

Review 9

Package golang-github-charmbracelet-x

NEW

#241

Package golang-github-aymanbagabas-go-udiff

NEW

#242

Package hey repack and QA

repack

#245

Package go-task

NEW

#169

Package ruby-unicode-plot

GCES

#82

Package imgsizer

#243

Package gtts-token

#252

Package hcxdumpool

GCES-2024-1

waiting-migration 3

Package adodb (NMU upload)

bug delay-7-days new-upstream-version

#172

Package gnome-shell-extension-easyscreencast new upstream version 1.9.0

GCES-2024-1 upload-done

#250

Package ruby-web-console

GCES-2024-1 upload-done

#257 Tomorrow

Closed 172

Package gpp (ITA)

upload-done

#247

Package kind

NEW

#55

Package smbmap

#239

Package cava-alsa

#233

Package smbmap

#223

Package python-bracex

GCES-2024-1 new-comer upload-done

#221

Package golang-github-dominikbraun-graph

NEW

#232

Package golang-github-fatih-camelcase-prepare for release

NEW

#154

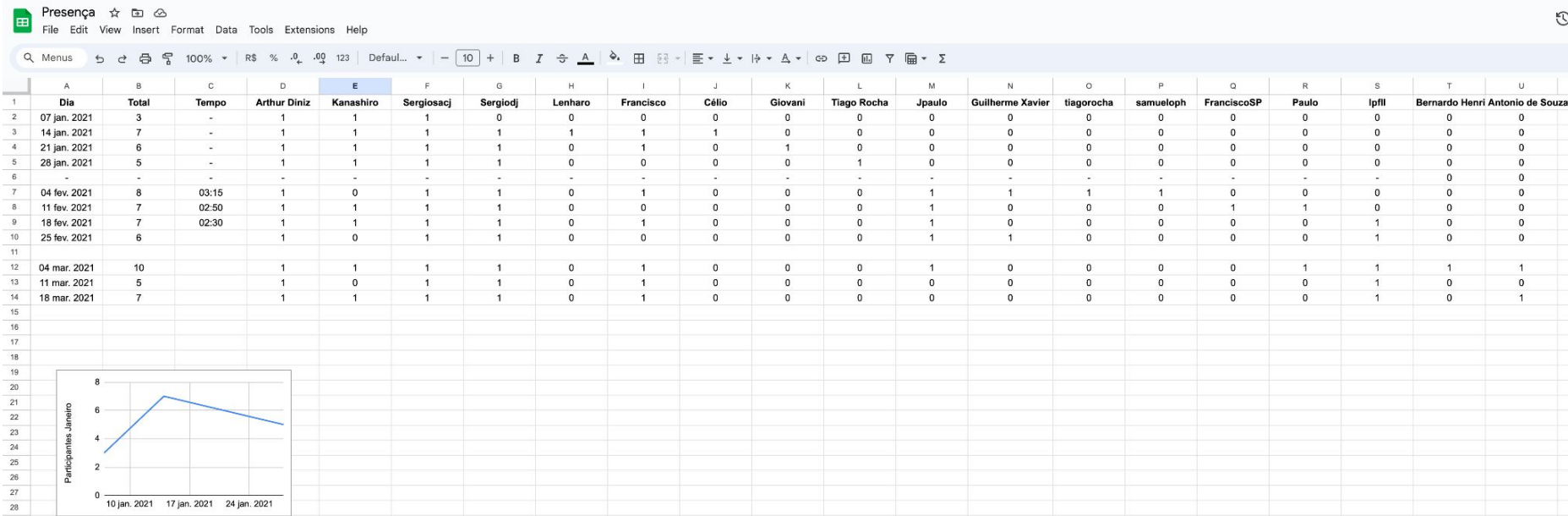
Package golang-github-liggitt-tabwriter

NEW

Questions

- How to know the health of the community?
- How to get insights?
- How to know if we are making the right decisions?
- How to track what people are doing?

First Attempt



Meet

meet.debianbsb.org

Reunião Debian Brasília

Inserir os dados abaixo é **opcional**, eles **não** serão compartilhados publicamente e servem apenas para medir o engajamento de reuniões da comunidade **Debian Brasília**.

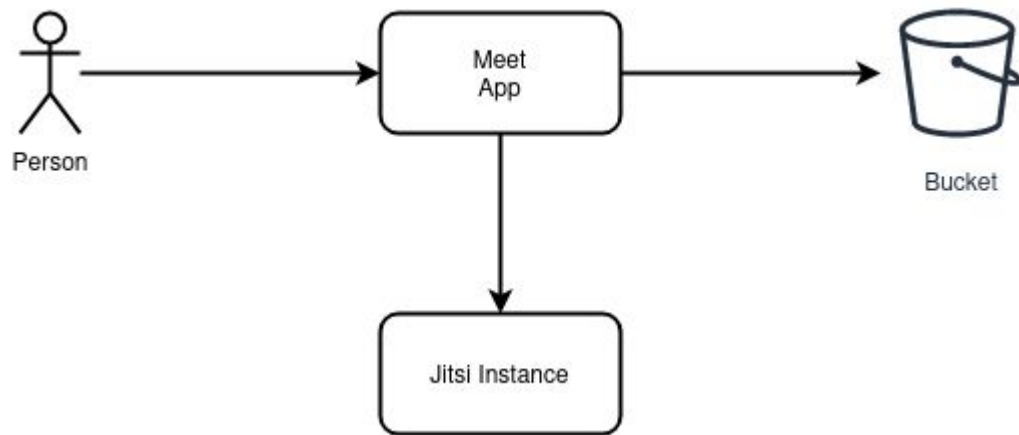
Antes de preencher, pedimos que se possível coloque o mesmo **Nome e Email** de contribuinte caso já seja um. :)

Nome (opcional)

Email (opcional)

ENTRAR

Meet




Meet Payload

```
1  {
2  |   "name": "Arthur Diniz",
3  |   "email": "arthurbdiniz@gmail.com",
4  |   "method": "POST",
5  |   "url": "http://meet.debianbsb.org/",
6  |   "headers": {
7  |       "user-agent": "Mozilla/5.0 (X11; Linux x86_64; rv:109.0) Gecko/20100101 Firefox/115.0",
8  |       "accept-language": "en-US,en;q=0.5"
9  |   },
10 |   "created_at": "2024-07-15T18:10:52.853241+00:00"
11 | }
```

Chat

The screenshot shows a Matrix chat window with a dark theme. On the left is a sidebar with navigation options: Convoos, Conversations, #debian-bsb (selected), arthurbdiniz@gmail.com, Notifications, Search, Add conversation, Connections, Account, Files, Help, and Log out. The main chat area displays a conversation in the #debian-bsb channel. The channel header includes the name, a description, and a participant count of 63. The chat history shows several messages from users MylenaTelegram, LeandroTelegram, SamuelHenriqueTelegram, LucasKanashiroTelegram, and CharlesMelaraTelegram. A quote from ga.debian.org is visible. The current message from arthurbdiniz asks, "What is on your mind, arthurbdiniz?". On the right, a participants list shows 63 members, including adrianorg, aquilamacedo, arthurbdiniz, athos, Charles, charlz, debianbsb, and others.

Convoos 🔍 🔔

#debian-bsb Canal Matrix: #debian-bsb.matrix.debian.social | Reunião Semanal Quintas às 19h no <https://meet.debi...>  **Participants (63)**

atualiza este repositório no salsa ou se quiser deixar comigo também.

“ga.debian.org/cgi-bin/vcswatch?package=libphp-adodb

feito, valeu!

MylenaTelegram[m] 21:50
Pessoal! Uma dúvida. Quando vou checar a Debian policy eu olho algum arquivo específico ?

LeandroTelegram[m] 21:51
SamuelHenriqueTelegram[m]: Obrigadol!

LucasKanashiroTelegram[m] 21:51
<**CharlesMelaraTelegram[m]**> "@lucaskanashiro do we still have..." <- maybe, I need to check. If I have one here I will bring it to you @weepingclown13 :)

LeandroTelegram[m] 21:52
MylenaTelegram[m]: tem um link, aguarde um momento que eu mando aqui www.debian.org/doc/debian-policy/
Debian Policy Manual — Debian Policy Manual v4.7.0.0
enviado



LucasKanashiroTelegram[m] 21:52
LucasKanashiroTelegram[m]: ah just saw puida's reply 😊 lucky you

MylenaTelegram[m] 21:52
LeandroTelegram[m]: Obrigada

SamuelHenriqueTelegram[m] 21:53
<**LeonardoGonavesMachadoTelega**> "image.jpeg" <- É esse mesmo, mas você terá que instalar o pacote, daí o comando irá atualizar os arquivos naquele diretório, e você copia eles para o empacotamento, faz sentido?

weepingclownTelegram[m] 21:53
LucasKanashiroTelegram[m]: forget you saw that, I was counting both as separate :)

LeandroTelegram[m] 21:53

What is on your mind, arthurbdiniz?  

Participants (63)

Members

- adrianorg
- aquilamacedo
- arthurbdiniz
- arthurbdiniz[m]1
- athos
- athos[m]
- Charles[m]
- charlz
- debianbsb
- debianuser93
- felipegmaia4191
- FranciscoPena[m]
- gitlab[m]
- hiagofranco[m]
- HookshotBot[m]
- jesualva-bot
- Jesualva[m]1
- jiande2020
- JoaaoNobrega[m]
- JoaoPedro
- JoaoPedroNobrega[m]
- johnson_dawson[m]1
- kanashiro
- kanashiro[m]1

Chat

```
2024-03-01T00:32:30 0 -gitlab[m]- [debian-brasilia-team/docs] puida commented on issue #123: Package dasel
2024-03-01T00:32:30 0 -gitlab[m]-
2024-03-01T00:32:30 0 -gitlab[m]- > Marked the MR as draft for now while I solve the issues that we discussed in the meeting today.
2024-03-01T00:34:25 0 <lucascastro> SrgioCiprianoTelegram[m]: neste endereco https://bolha.video/debian-bsb |
2024-03-01T00:34:53 0 <lucascastro> Não há ninguém.
2024-03-01T00:35:13 0 <Leandro[m]> jitsi.debian.social voltou
2024-03-01T00:35:24 0 <Leandro[m]> agora já acabou também
2024-03-01T00:35:29 0 <lucascastro> hahaa
2024-03-01T00:35:31 0 <lucascastro> de boa.
2024-03-01T00:35:35 0 <lucascastro> mas qual era o link?
2024-03-01T00:36:03 0 <Leandro[m]> lucascastro: https://jitsi.debian.social/debianbrasilia
```

Requirements

- Stack in **Python**
- Fully **Open Source**
- No real time data
- Can contain full snapshots of data
- Extensible to add more data sources
- Exported in a web **interface** with easy access
- **Full Load** and **Incremental Load**

The ETL Process Explained



Extract

Retrieves and verifies data
from various sources



Transform

Processes and organizes
extracted data so it is usable



Load

Moves transformed data
to a data repository

Extract

- Define the data sources
 - Meet
 - Chat (Convos)
 - Salsa Issues and comments
 - UDD
- Format
 - Log
 - JSON
 - SQL
- Create automation
 - Cron scripts
 - Streaming
- Save raw data
 - Bucket
 - Database

Bucket

User

Object Browser

Access Keys

Documentation

Administrator

Buckets

Policies

Identity





Monitoring

Events

Tiering

Object Browser

Filter Buckets

Name	Objects	Size
 convos	11	601.8 KiB
 meet	468	381.8 KiB
 salsa	4,499	7.3 MiB
 udd	342	376.2 KiB

Transform

- Remove empty fields
- Flatten nested object
- Remove Personally Identifiable Information (PII)

Load

- Use a table format to catalog and organize data
 - Apache Iceberg
 - Apache Hudi
 - Delta Lake
- Load into a warehouse

Apache Iceberg

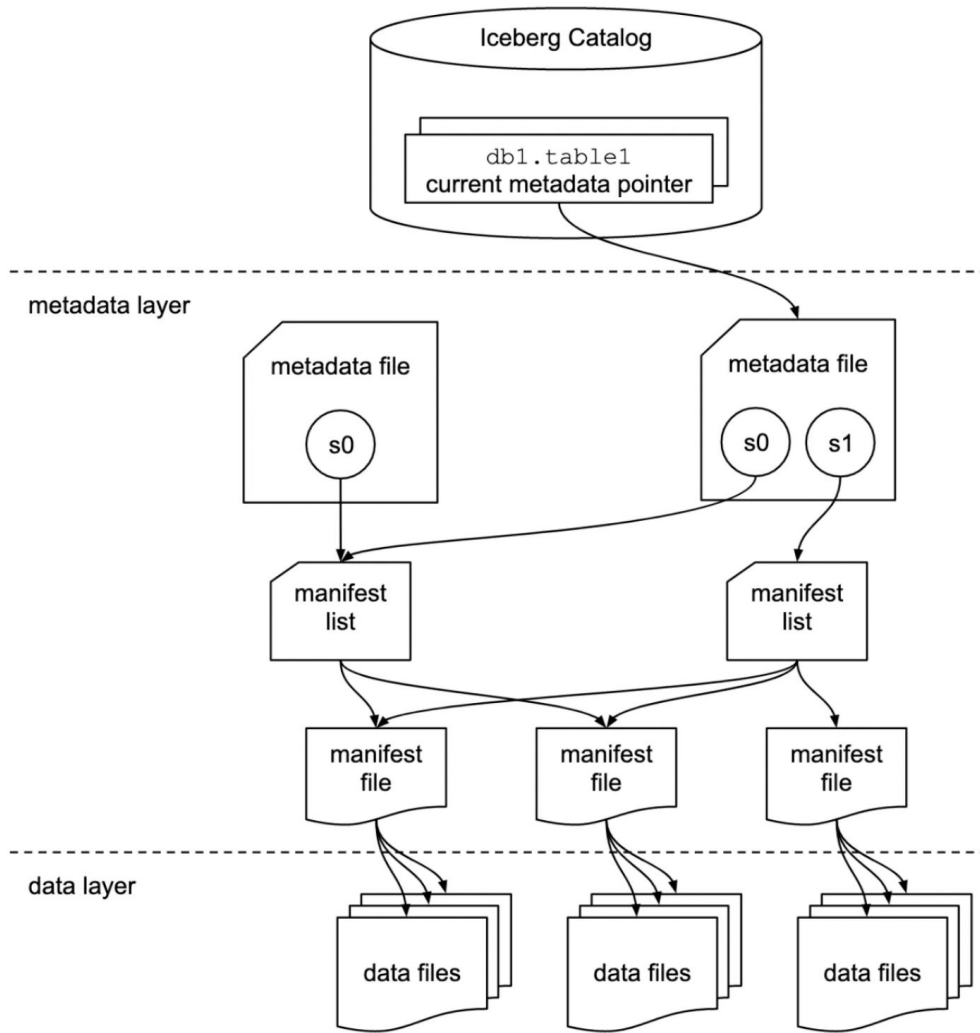
The open table format for analytic datasets.

- Expressive SQL
- Full Schema Evolution
- Hidden Partitioning
- Time Travel and Rollback
- Data Compaction



Iceberg Architecture

1. The Iceberg catalog
2. The metadata layer
3. The data layer



Trino

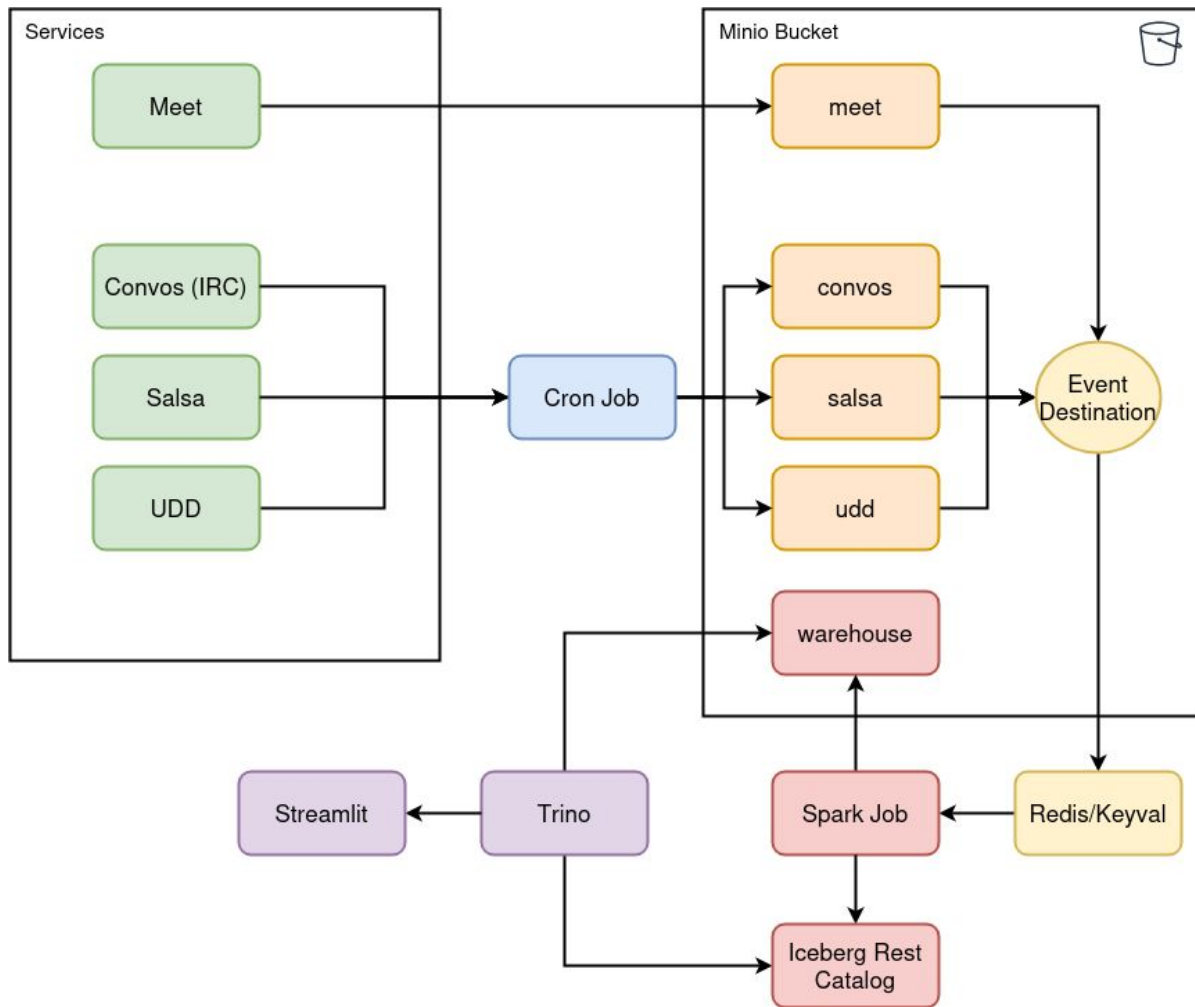
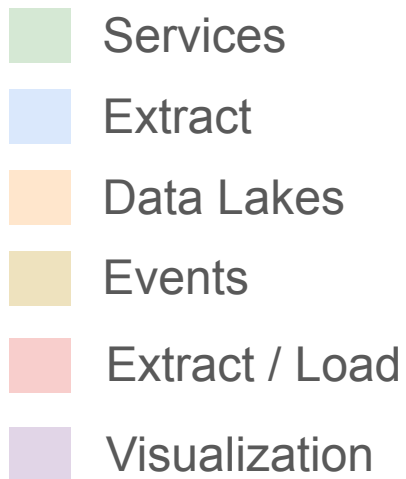
Trino is an SQL query engine, that works with BI tools such as R, Tableau, Power BI, Superset and many others.

```
cur.execute(  
    f"""  
    SELECT *  
    FROM db.udd  
    WHERE metadata_year = {last_entry_year}  
    AND metadata_month = {last_entry_month}  
    AND metadata_day = {last_entry_day}  
    """  
)  
rows = cur.fetchall()
```



trino

Data Flow



Demo

Next steps

- Export processed data to contributors.d.o
- Add more relevant data
- Make service available to all
- Find more use cases for the data

Thank You