GitHub Archive ETL for Sparrow Analytics

Team MINIRO

GitHub Archive Data

a project to **record** the public GitHub timeline, **archive it**, and **make it easily accessible** for further analysis

Event Examples

- Push
  commits are pushed to a repository branch or tag.
- Gollum
  human-readable files created or updated
- Create
  a git branch or is created
- Pull Request
  a contributor requests permission to merge a protected branch

Data Lake

Azure Data Lake Storage Gen2 = ADLS Gen1 + Blob storage

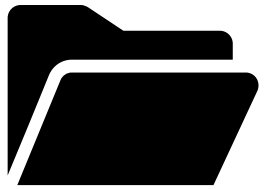
- hierarchical namespace, file-level security
- Blob storage

→ file system performance at object storage scale and prices

File Structure

LANDING ZONE

Compressed JSON format
Raw GitHub Archives



BRONZE LAYER

Parquet format
Partitioned by day

SILVER LAYER

Cleaned data
Snowflake structure

GOLD LAYER

Aggregated data
Ready for data warehouse

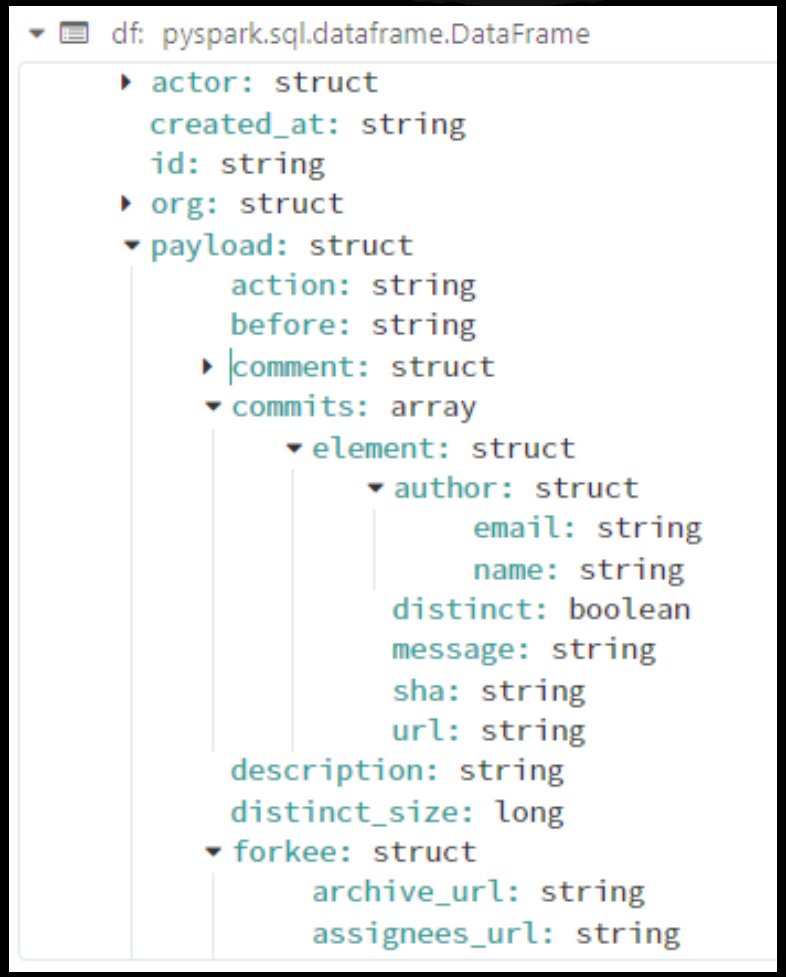
Cleaning

Goal: Process the data so that it is in a flattened tabular format

- Keep as much data as possible
  - Futureproof for any new business usages
- Drop any unnecessary columns
  - Columns that are all null, columns that are a url or contain data on 'gravatar' (globally recognized avatar)

Flattening the data

Issue: Nested Values



Solution:
Unnest the JSON structs into separate columns

Issue: JSON data is Flexible, Tables are not

```
{
  "id": 34,
  "type": "car",
  "make": "Honda",
  "opts": {
          "seats": 5,
          "airbags": 0,
          "ac": True
  }
}
```

```
{
  "id": 87,
  "type”: "motorbike",
  "make": "Ducati",
  "opts": {
        "sidecar": True,
        "transmission": "chain"
  }
}
```

Vehicle DataFrame							
id	type	make	**opts_seats**	**opts_airbags**	**opts_ac**	**opts_sidecar**	**opts_windshield**
34	car	honda	5	0	true	NULL	NULL
87	motorbike	ducati	NULL	NULL	NULL	True	chain

Flattening the data

Issue: Event Types contain only some payload columns

payload_distinct_size	payload_forkee	payload_head	payload_issue
null	null	null	null
null	null	null	null
1	null	a76e751b921a74f0c58f0c5f680c4415d88f04fd	null
null	null	null	null
1	null	d9d3b02535f14610efdbd9a3e543d7318d34e774	null
null	null	null	▸ {"assignee": null, "assignees": [], "body": "This was crea "comments_url": "https://api.github.com/repos/SouthAf "2017-01-01T18:45:18Z", "events_url": "https://api.github "html_url": "https://github.com/SouthAfricaDigitalScienc "default": false, "id": 308743652, "name": "build", "url": " deploy/labels/build"}, {"color": "ededed", "default": false, "https://api.github.com/repos/SouthAfricaDigitalScience "https://api.github.com/repos/SouthAfricaDigitalScience null, "number": 123, "pull_request": null, "repository_url"

Solution:

> Separate the data into 14 different Event Type tables

> Flatten one layer only

> Drop null columns

> Flatten completely

Two Approaches to Flatten

Hard Coding

Pros:

- ✓ Explicit about the kept columns
- ✓ Faster run time

Cons:

- Manually selecting each column for every flattened layer
- Needs to be changed if the schema of the data changes

Automating

Pros:

- ✓ Reusable code for data with JSON type columns
- ✓ Data does not have to be manually explored for each flattened layer
- ✓ Templates can be used to fix the issue of changing data schemas

Cons:

- Slower run time
- More difficult to debug

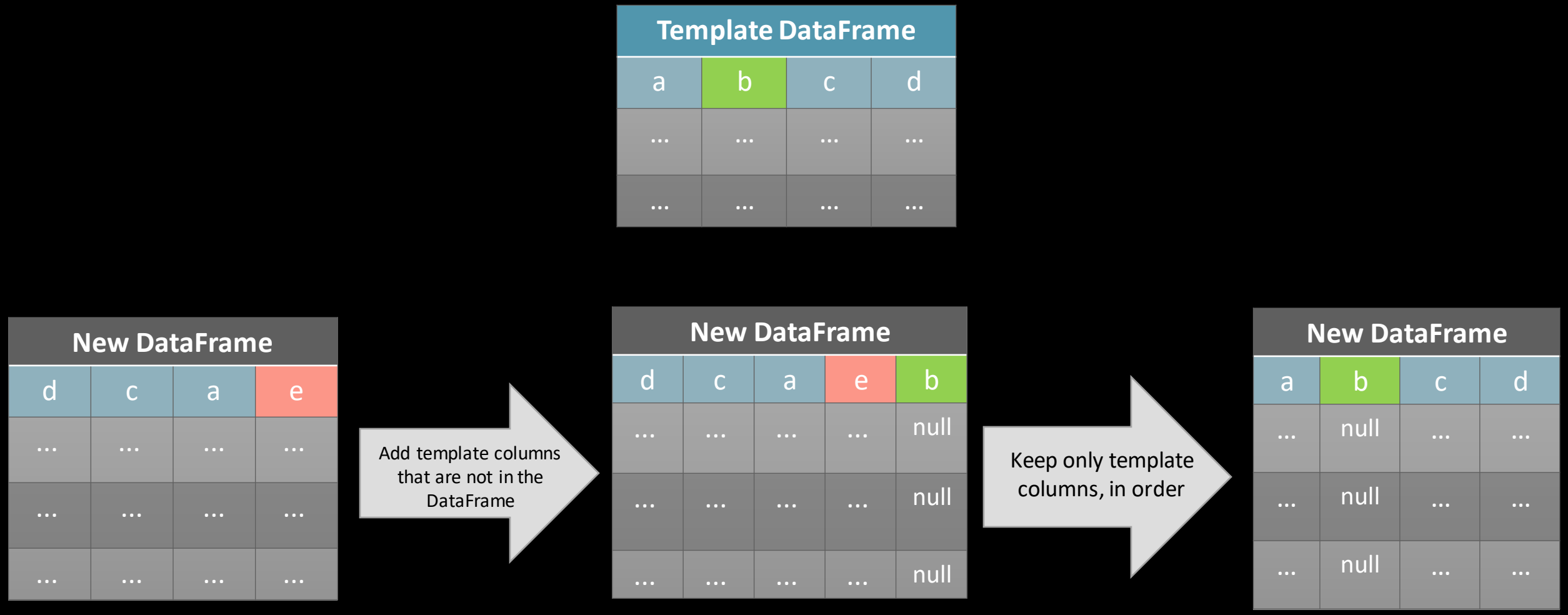
Issue: Data Schema Changes

The data added....

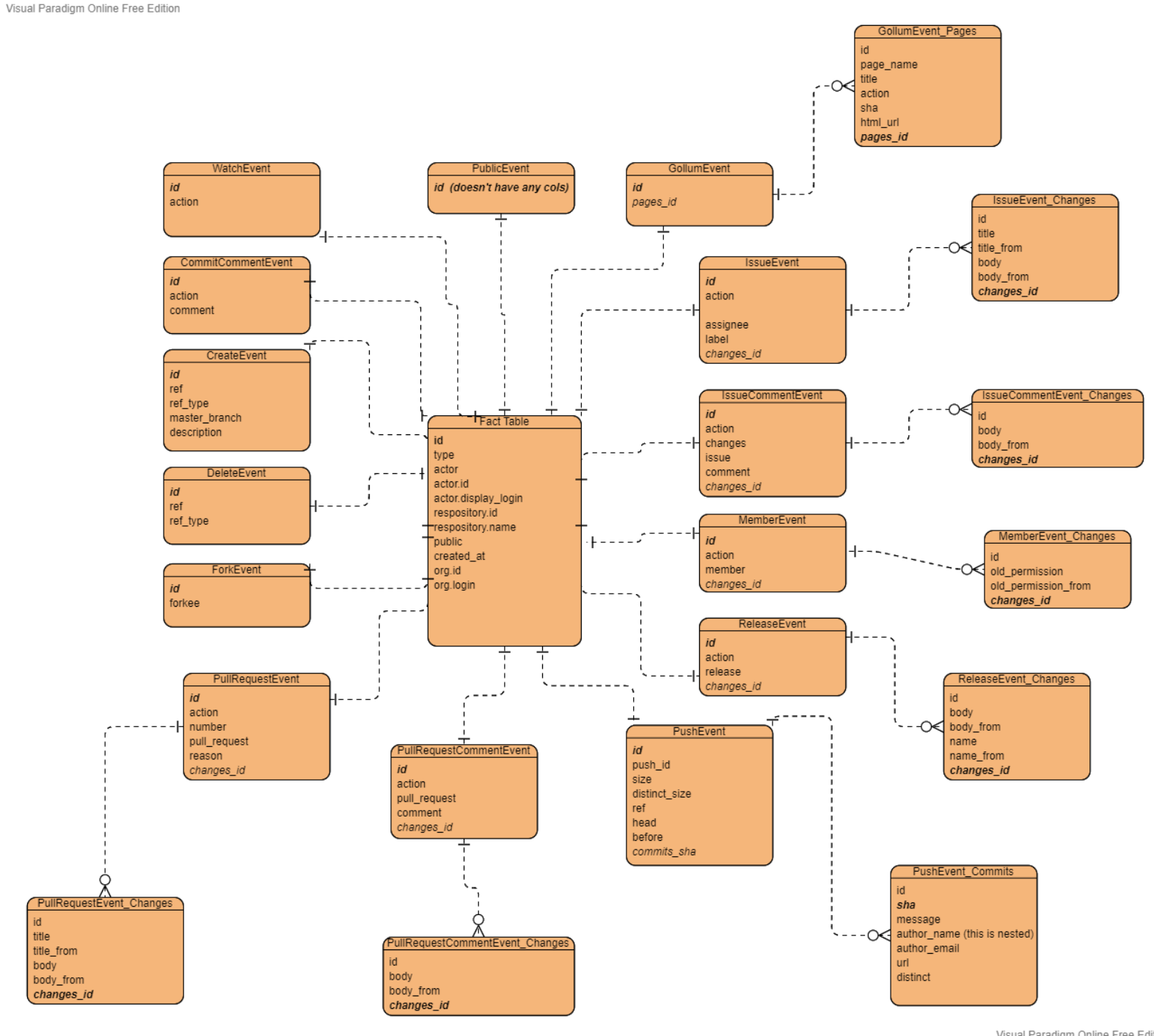
- contains extra columns
  - Any extra columns need to be dropped

- does not contain the columns from the current schema
  - Add null columns if data doesn't contain a column that is in the template

Solution: Templates Ensure Integrity



ERD


Visual Paradigm Online Free Edition
GollumEvent_Pages
id
page_name
title
action
sha
html_url
pages_id
WatchEvent
id
action
PublicEvent
id (doesn't have any cols)
GollumEvent
id
pages_id
IssueEvent_Changes
id
title
title_from
body
body_from
changes_id
CommitCommentEvent
id
action
comment
IssueEvent
id
action
assignee
label
changes_id
CreateEvent
id
ref
ref_type
master_branch
description
Fact Table
id
type
actor
actor.id
actor.display_login
respository.id
respository.name
public
created_at
org.id
org.login
IssueCommentEvent
id
action
changes
issue
comment
changes_id
IssueCommentEvent_Changes
id
body
body_from
changes_id
DeleteEvent
id
ref
ref_type
MemberEvent
id
action
member
changes_id
MemberEvent_Changes
id
old_permission
old_permission_from
changes_id
ForkEvent
id
forkee
ReleaseEvent
id
action
release
changes_id
PullRequestEvent
id
action
number
pull_request
reason
changes_id
ReleaseEvent_Changes
id
body
body_from
name
name_from
changes_id
PullRequestCommentEvent
id
action
pull_request
comment
changes_id
PushEvent
id
push_id
size
distinct_size
ref
head
before
commits_sha
PullRequestEvent_Changes
id
title
title_from
body
body_from
changes_id
PullRequestCommentEvent_Changes
id
body
body_from
changes_id
PushEvent_Commits
id
sha
message
author_name (this is nested)
author_email
url
distinct
Visual Paradigm Online Free Edition