So You Inherited A Database

Corey Huinker

Corlogic

PgConf NYC 2021

Who Am I

- database programmer, consultant
- NYC based
- occasional postgres contributor

Software Evolves Like The Earth

https://en.wikipedia.org/wiki/File:The_Earth_seen_from_Apollo_17.jpg

Prelude - Hadean Eon

- In The Beginning, there was the Green Field
- And a small development team of developers

https://commons.wikimedia.org/wiki/File:Hadean.png

Late Heavy Bombardment

- Fast Paced Environment
- Requirements Change Daily
- Requirements rarely documented

Key Code Features of Late Heavy Bombardment

- code comments focused primarily on the author's own mental state
- commits < 10 minutes apart for stretches of 15 hours or more
- commit messages < 3 words

MVP - The Archaen Eon

- Things Seem To Work
- People Relax For A Bit (Photosynthesis)

https://en.wikipedia.org/wiki/Archean#/media/File:Archean.png

Proterozoic

- Ideas for How To Expand (Eukaryotes)
- New Business Partnerships (multicellular life)
- Competition Emerges (other multicellular life)

https://en.wikipedia.org/wiki/File:Life_in_the_Ediacaran_sea.jpg

Key Features Of The Proterozoic Era

- Bolt-on features added to add green checkboxes to someone's slide deck
- A realization that relaxing was unwarranted

Production Maintenance

(Cambrian Explosion)

- Unforseen Scaling Issues Arise
- Reality often does not fit the data model
- Rushed design decisions
- No more permanent solution than a temporary fix

https://en.wikipedia.org/wiki/Cambrian#/media/File:Graphic_kingdom_growth_series.jpg

Frustration

(Cambrian-Ordovician-Silurian Extinctions)

- Developers accustomed to greenfield development
- Maintenance is a drag
- Hire more developers (in their own image)...
- ...who mostly want to do greenfield development.
- And leave when it is no fun anymore.

Early Devonian

- New people attempt to understand the existing systems
- With no documentation
- And the authors are gone or checked out
- External demands for new features have not slowed
- Frustration Mounts
- Decide it's easier to replace the app
- But we can't migrate the database so it stays the same

https://en.wikipedia.org/wiki/Ammonoidea#/media/File:Pleuroceras_solare,_Little_Switzerland,_Bavaria,_Germany.jpg

Devonian Peak (Land!)

- New developer writes new application to replace the old application
- That does 80% of what the old application did.
- But the last 20% is hard.

https://en.wikipedia.org/wiki/Devonian#/media/File:Devonianscene-green.jpg

Devonian Extinction

Developer has options:

1. Invest hard work in understanding and replicating the last 20%
2. Go into management and hire the developer to do the hard work
3. Make a slide deck and blog post which are used to get a new job

Carboniferous (Amphibians)

- 2 systems to support
- Interactions with the database are often contradictory
- Neither system is documented
- Both have maintenance issues

https://en.wikipedia.org/wiki/Carboniferous#/media/File:Pederpes22small.jpg

Permian Extinction

Developer has options:

1. Invest hard work in understanding two applications and how each is only an 80% solution.
2. Scrap both of them and start over.
3. Listen to Application Vendor Sales Pitches
4. Update LinkedIn

https://en.wikipedia.org/wiki/Dimetrodon#/media/File:Dimetrodon_incisivum_01.jpg

Triassic - The Vendor Solution

- The Vendor Software does 80% of what is needed
- The original 2 systems must be maintained
- Now you have 3 problems

https://en.wikipedia.org/wiki/Triassic#/media/File:Petrified_Forest_National_Park-Rainbow_Forest_Museum-1.jpg

Jurrassic - Interconnects

- Interconnections between in-house and vendor solutions are complex and buggy
- Promises of zero-code solutions fail to mention extensive manual configuration
- What has stayed the same through all of this?

https://en.wikipedia.org/wiki/Jurassic#/media/File:Monolophosaurus_jiangi.jpg

Cretaceous - The Cloud

- Vendor Solution With Hosting
- 4 problems

https://en.wikipedia.org/wiki/Cretaceous#/media/File:9182_-_Milano_-_Museo_storia_naturale_-_Derasmosaurus_pietraroiae_-_Foto_Giovanni_Dall'Orto_22-Apr-2007_(cropped).jpg

Cenozoic

Your first day

https://en.wikipedia.org/wiki/Cave_painting#/media/File:Bestias11.JPG

Your Database Is More Than Just The Server

- the databases
- the backups
- the input sources
- the customers
- the SLAs

Step 0 - The Backups

- Are they?
- *Where* are they?
- How many are there?
- When is the last time a restore was tested?
- How can you test a restore right now?

Define the SLAs

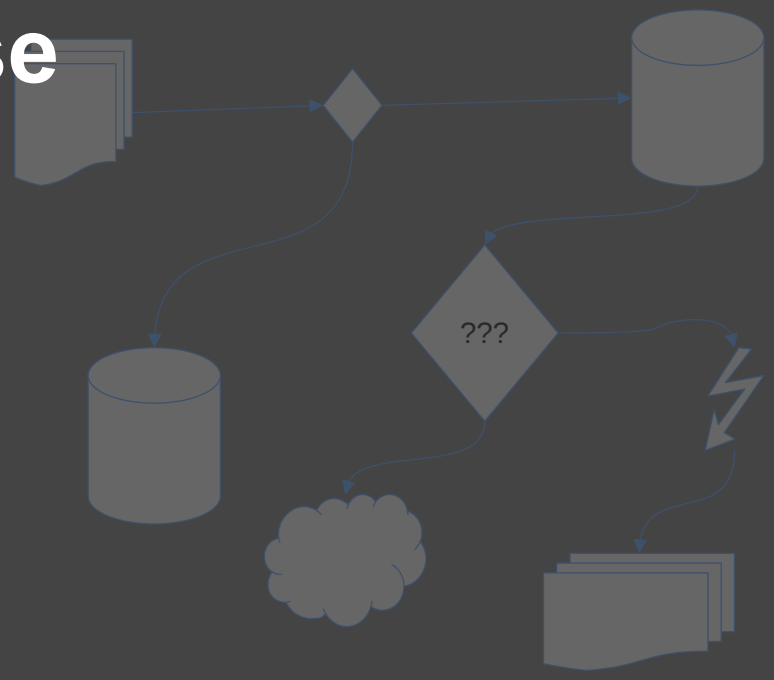
- What's an SLA?
- Written or Unwritten
- Find the Gaps
- This is your wiggle room

Get a Maintenance Window

- What do apps do when the DB is offline?
- Not all maintenance is planned
- Physical vs Logical / Partial offline
- Have a plan for each application

Document Your Database

- Is there a data dictionary?
- Create one
- Tables
- Columns
- Constraints
- Indexes
- Internal vs External



Document Your Database - Internal

- `pg_description`
- Automatic audits possible
- Poor outside visibility

Document Your Database - External

- Spreadsheets, documents, etc
- Good visibility
- Automatic audits impossible

Document Your Tables

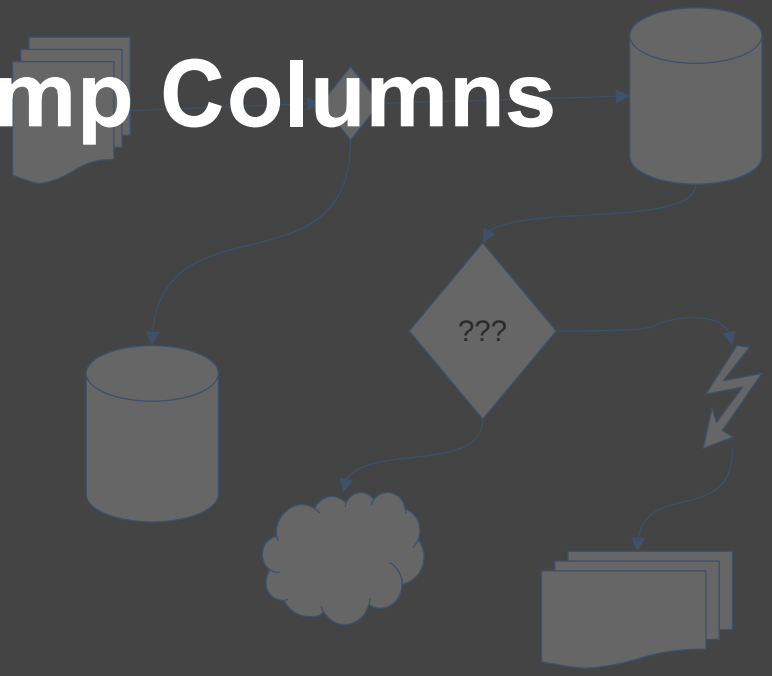
- Why is this table?
- How old is the data in it?
- What apps use it?
- Where does input data come from?
- What are the consumers of this data?
- Who can see this table?
- Is there a gap between those two?

Document Your Columns

- Is this column populated?
- Is it typed correctly?
- Is it named correctly?
- What does a NULL value mean?
- Can you limit valid values?
- Do existing values meet those restrictions?
- Is it PII?
- Is it PII in conjunction when combined with other columns?

Document Your Timestamp Columns

- Are they timezone aware?
- Do they imply a range (begin/end)?
- Are there collisions or overlaps?
- When is "never"?
- When is "always"?
- Should this have been a date?



Document Your Date Columns

- What date where?
- Are collisions ok?
- Should this have been a timestamp?

Document Your Numberlike Columns

- Do they have the proper precision?
- Are they over-promising?
- What does a negative value mean?

Document your document (JSON) columns

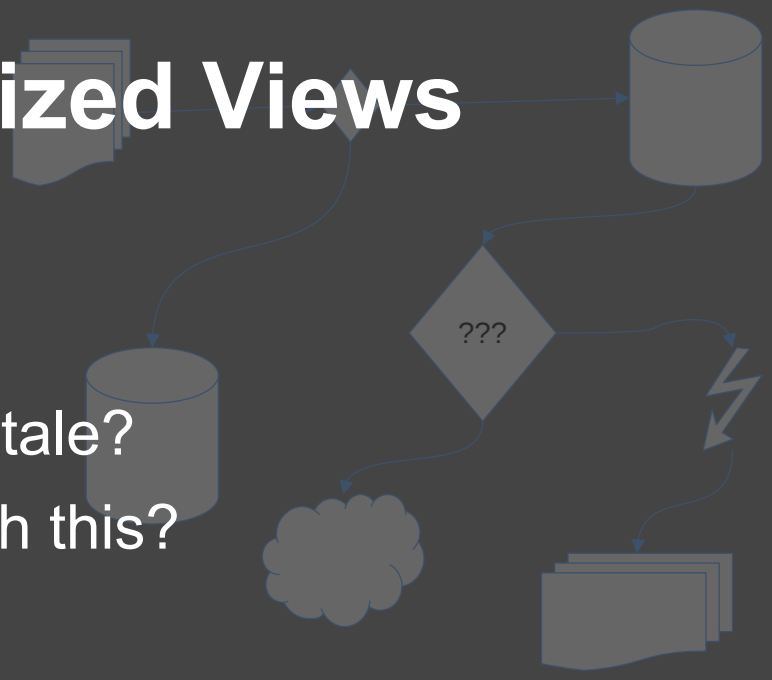
- Why isn't this several regular columns?
- Are we searching on elements within this document?
- Should we cherry-pick some attributes into columns?
- Perhaps candidates for `GENERATED ALWAYS AS...`

Document Your Views

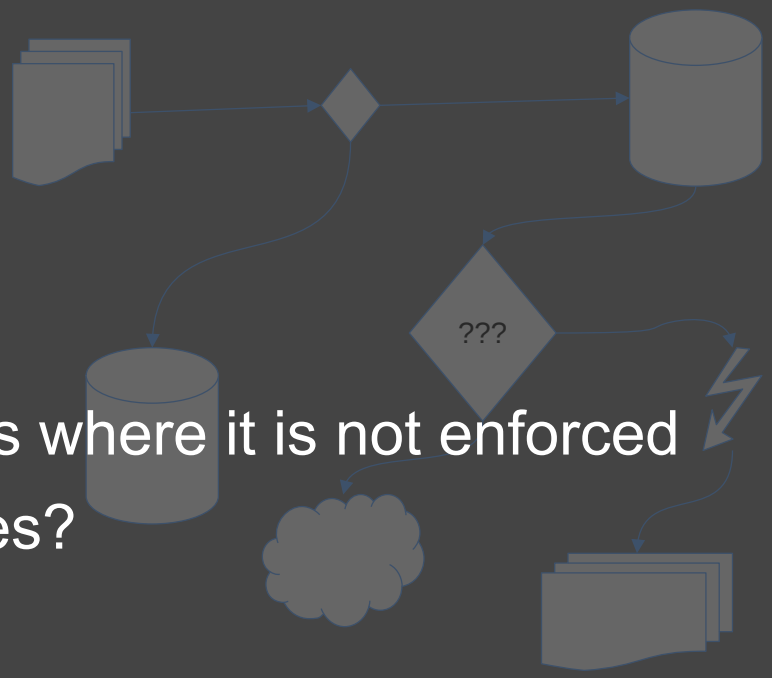
- What does this view simplify?
- Is it named correctly?
- Does it have any consumers?
- Is it a permission barrier?

Document Your Materialized Views

- Who are the consumers?
- How often is it refreshed?
- What happens when the data goes stale?
- Are the consumers aware and ok with this?
- Should this be a regular view?
- Should this be a regular table?



Document Uniqueness



- Apps sometimes assume uniqueness where it is not enforced
- Does a tiebreaker exist in those cases?

Document Constraints

- Is the contraint too narrow?
- Is the contraint too broad?
- Is it permanent or temporary?

Document Indexes

- Is it being used (`pg_stat_user_indexes`)?
- Has it been `REINDEX`ed recently?
- Could it be split into partial indexes?

Document Relationships

- Are relationships defined with foreign keys?
- What unenforced relationships exist?
- How do these relationships extend the privacy context?

Identify Data Sources

- Where does this data come from?
- Are you responsible for the source as well?
- Can that source be paused?
- Should that data source be under a privacy regimen?

Document Data Lifecycle

- When does this data arrive?
- When does the importance decline?
- When does its usefulness end?
- How should expired data be handled?
- Is there an easy way to identify expired data?
- Is there an easy way to isolate expired data?

Security

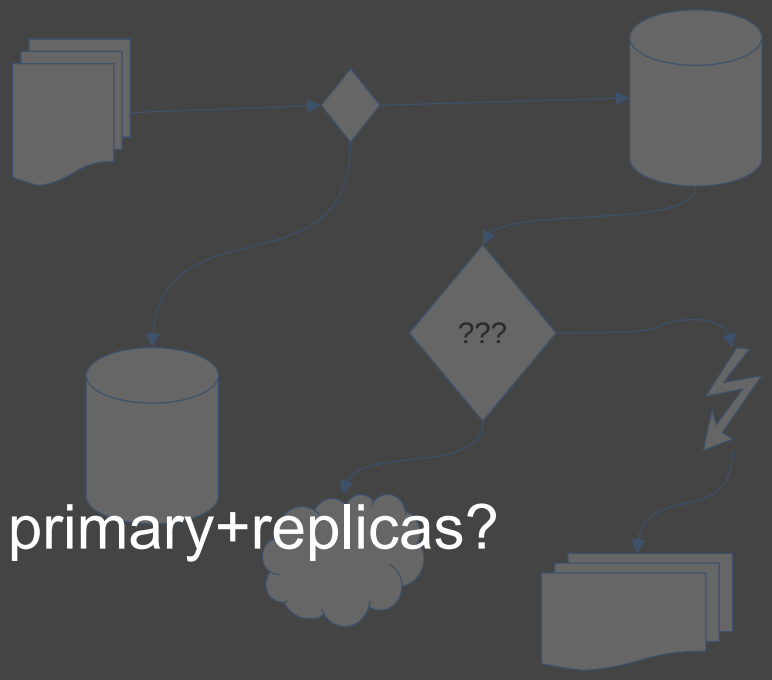
- How many roles exist in the system?
- Does each app have its own role?
- Does that role follow Principle of Least Privilege?
- Is there distinction between login roles and app roles?
- How often are credentials cycled?
- When was the last time each app tested credential cycling?

Security - Sad Reality

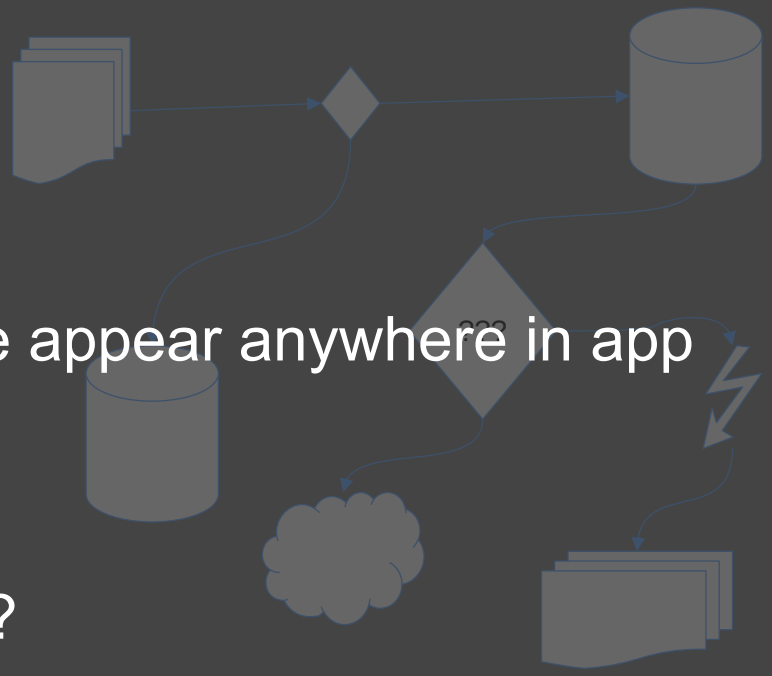
- ORMs
- Migrations
- Convenience
- Laziness

Failover

- Is there a replica?
- Can one machine handle the load of primary+replicas?



The Network



- Does the hostname of your database appear anywhere in app configuration?
- DNS is your friend
- When was the last DNS failover test?

The Downstream - Apps

- Is this app read only?
- What does this app do when it cannot connect?
- Can this app be paused?

The Downstream - Reports

- Where does this report land?
- Is there PII in the report?
- Does the landing zone have access controls?
- Does the landing zone have privacy controls?
- Does the landing zone have lifecycle policy?

Data Warehouse

- Does reporting happen in a separate database?
- Snapshots or Change Data Capture?
- What gets lost in translation?

Thank You