AUTOMARTS

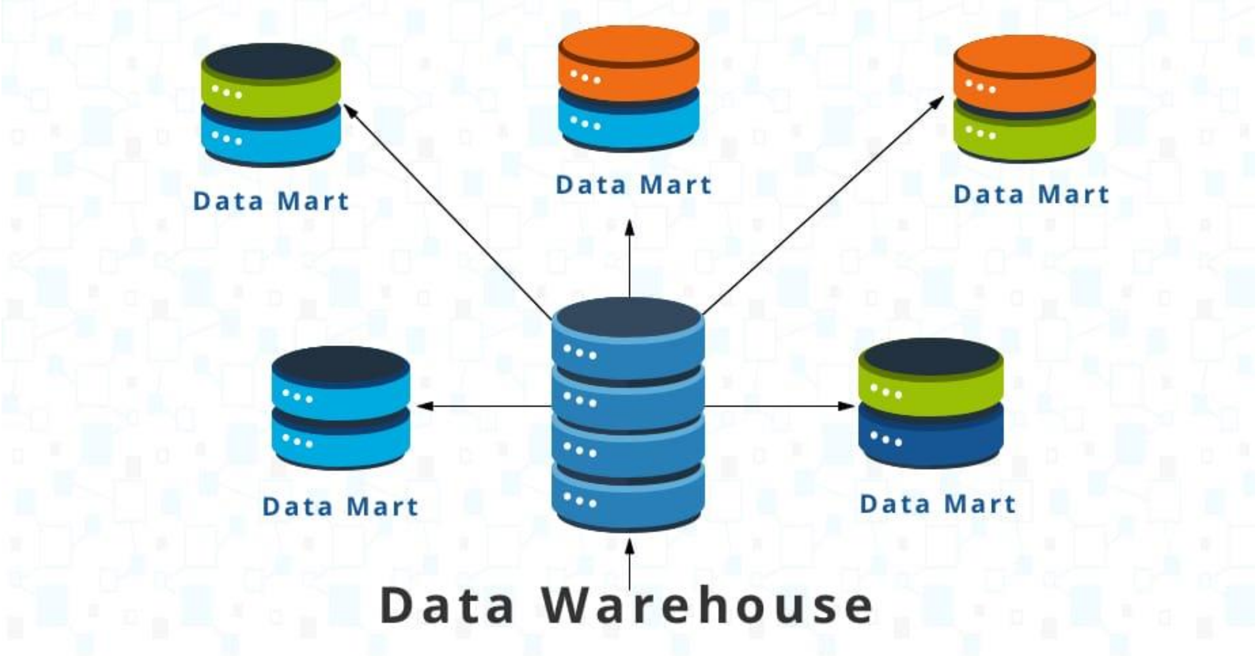
The Analyst-Owned ETL

Mayank Sardana
Christian Minich

Children's Hospital of Philadelphia®
Center for Healthcare Quality & Analytics

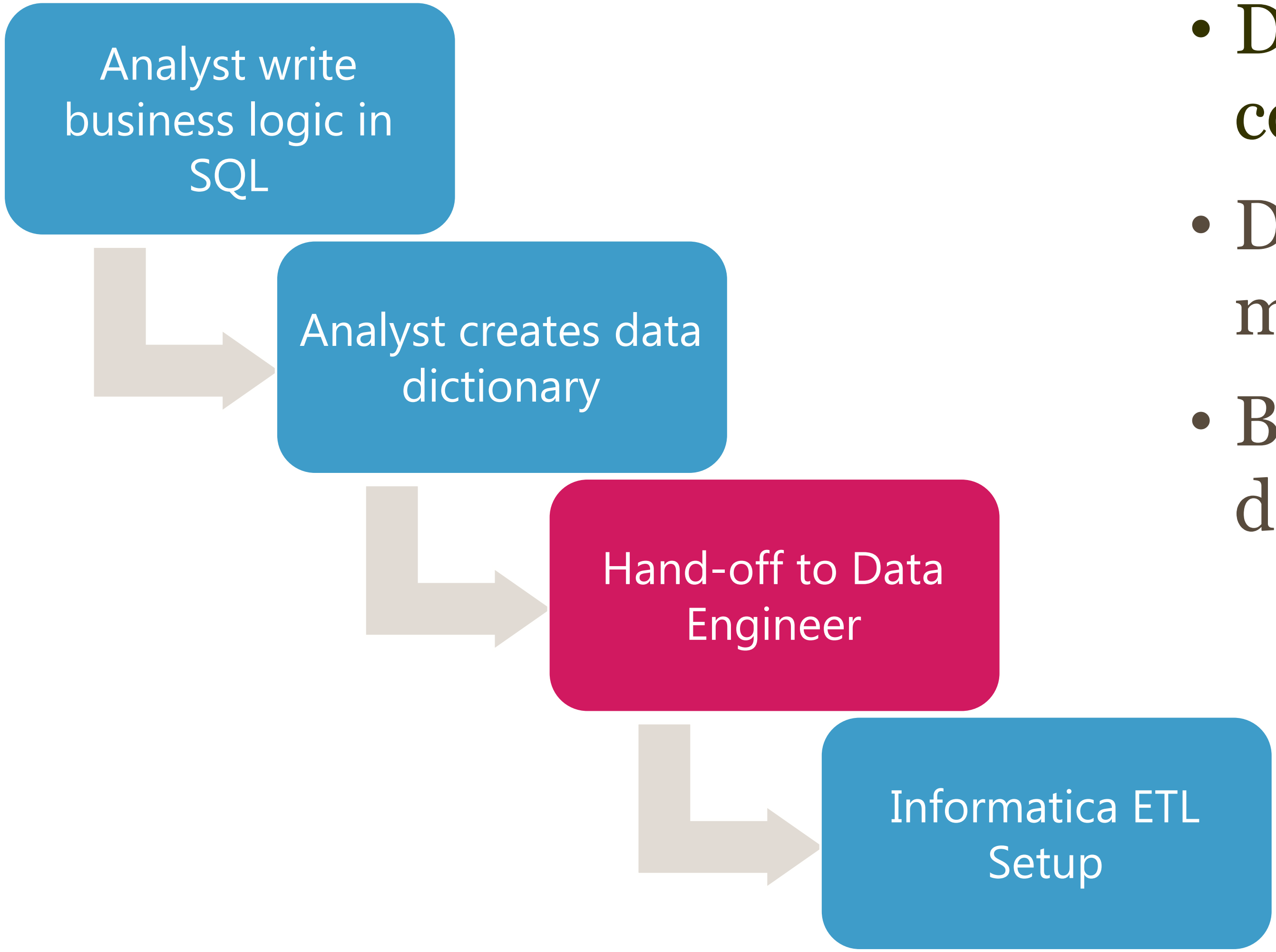
DATA MARTS

- Subject-specific data slices
- Supports constantly changing business definition and needs
- Logic typically goes in either BI tool, Views or ETL tool



Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

CREATING DATA MARTS • TRADITIONAL METHOD


Analyst write business logic in SQL

Analyst creates data dictionary

Hand-off to Data Engineer

Informatica ETL Setup

- Data engineers are not data content experts
- Data Analysts are not data modeling and ETL experts
- Both work in silo to create data marts

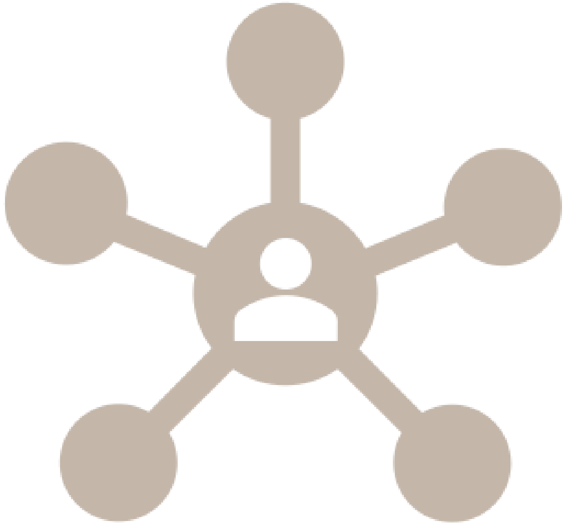
Children's Hospital of Philadelphia®
Center for Healthcare Quality & Analytics

ISSUES

- No code transparency
- Time consuming and error prone
- Reusability of ETL mappings
- Unsustainable for high Analyst to Engineer ratio (e.g. 10:1)
- Tedious to make changes to static ETL process
- Scalability

DESIGN PRINCIPLES • ANALYST OWNED ETL

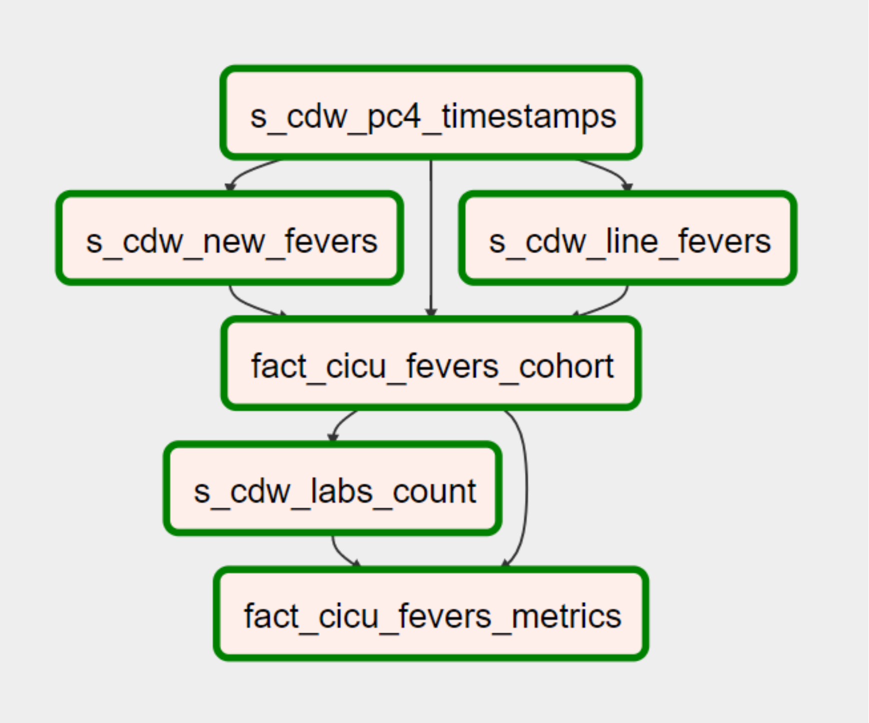
- Complete control to Analysts over their ETL
- GitHub as source of truth for any business logic
- Minimal/no Data Engineer involvement
- Minimal process administration



APACHE AIRFLOW OVERVIEW

- Programmatic ETL tool written in Python
- Workflows represented as Directed Acyclic Graph (DAG)
- Web interface!!!
- Extensive Operators for AWS S3, Python, Hive, Slack, Bash etc.
- Sensors for dependency management e.g. File Watcher, SQL sensor, DAG sensor
- Built-in scheduler and webserver

DAG



Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

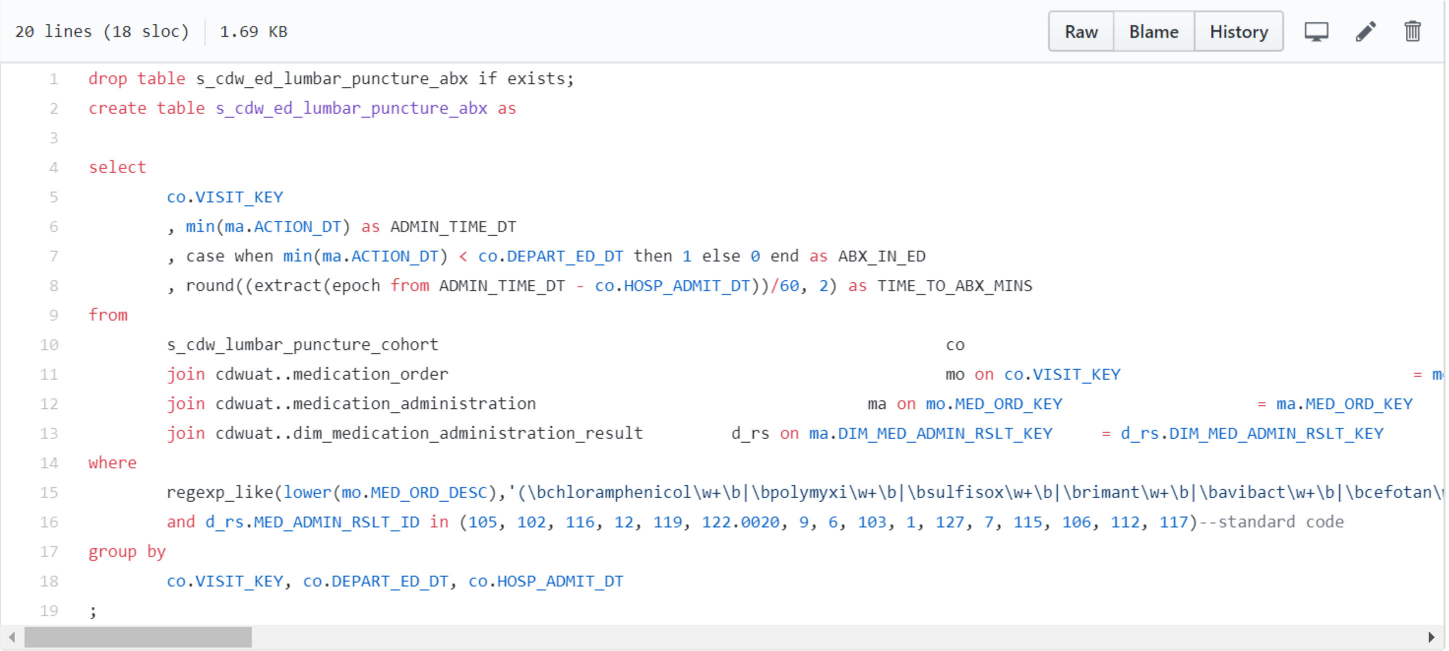
AUTOMARTS • INITIAL DESIGN

NON-PROD

AUTOMARTS • INITIAL DESIGN

NON-PROD

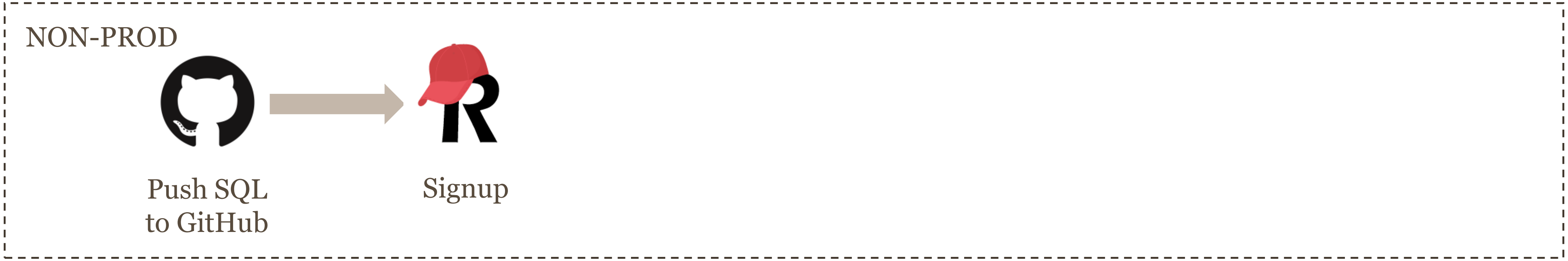
Push SQL
to GitHub



```sql
drop table s_cdw_ed_lumbar_puncture_abx if exists;
create table s_cdw_ed_lumbar_puncture_abx as

select
        co.VISIT_KEY
        , min(ma.ACTION_DT) as ADMIN_TIME_DT
        , case when min(ma.ACTION_DT) < co.DEPART_ED_DT then 1 else 0 end as ABX_IN_ED
        , round((extract(epoch from ADMIN_TIME_DT - co.HOSP_ADMIT_DT))/60, 2) as TIME_TO_ABX_MINS
from
        s_cdw_lumbar_puncture_cohort                                                    co
        join cdwuat..medication_order                                                   mo on co.VISIT_KEY                               = m
        join cdwuat..medication_administration                                  ma on mo.MED_ORD_KEY                     = ma.MED_ORD_KEY
        join cdwuat..dim_medication_administration_result         d_rs on ma.DIM_MED_ADMIN_RSLT_KEY      = d_rs.DIM_MED_ADMIN_RSLT_KEY
where
        regexp_like(lower(mo.MED_ORD_DESC),'(\bchloramphenicol\w+\b|\bpolymyxi\w+\b|\bsulfisox\w+\b|\brimant\w+\b|\bavibact\w+\b|\bcefotan\
        and d_rs.MED_ADMIN_RSLT_ID in (105, 102, 116, 12, 119, 122.0020, 9, 6, 103, 1, 127, 7, 115, 106, 112, 117)--standard code
group by
        co.VISIT_KEY, co.DEPART_ED_DT, co.HOSP_ADMIT_DT
;
```

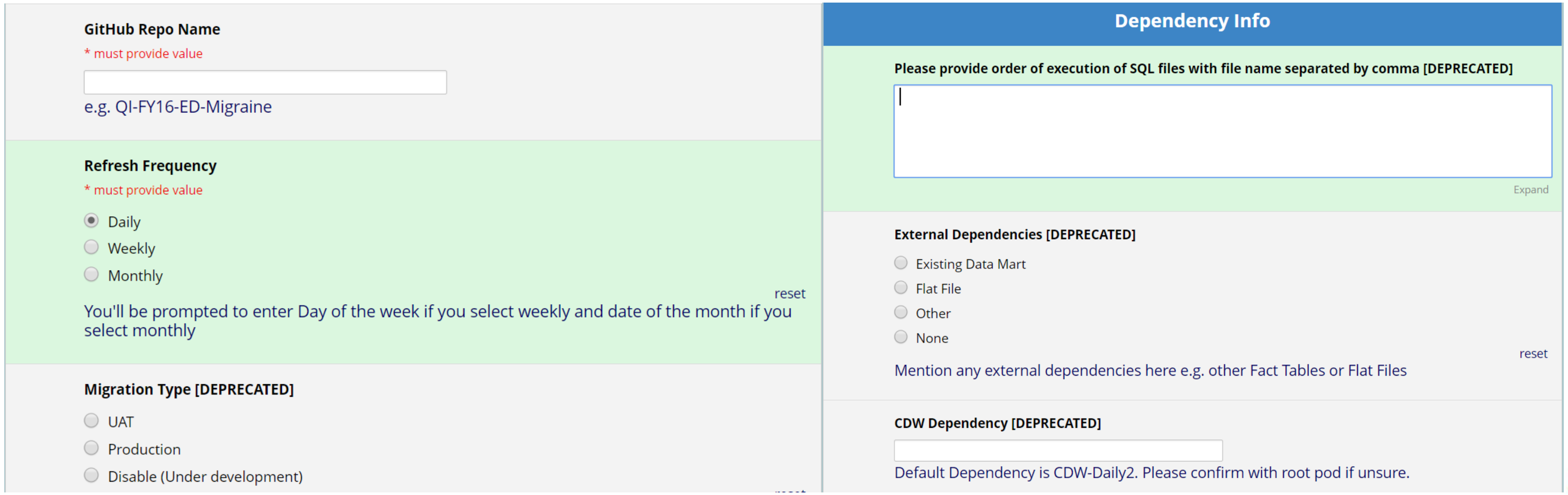
AUTOMARTS • INITIAL DESIGN



NON-PROD

Push SQL to GitHub

Signup



GitHub Repo Name

* must provide value

e.g. QI-FY16-ED-Migraine

Refresh Frequency

* must provide value

● Daily
○ Weekly
○ Monthly

You'll be prompted to enter Day of the week if you select weekly and date of the month if you select monthly

Migration Type [DEPRECATED]

○ UAT
○ Production
○ Disable (Under development)

Dependency Info

Please provide order of execution of SQL files with file name separated by comma [DEPRECATED]

External Dependencies [DEPRECATED]

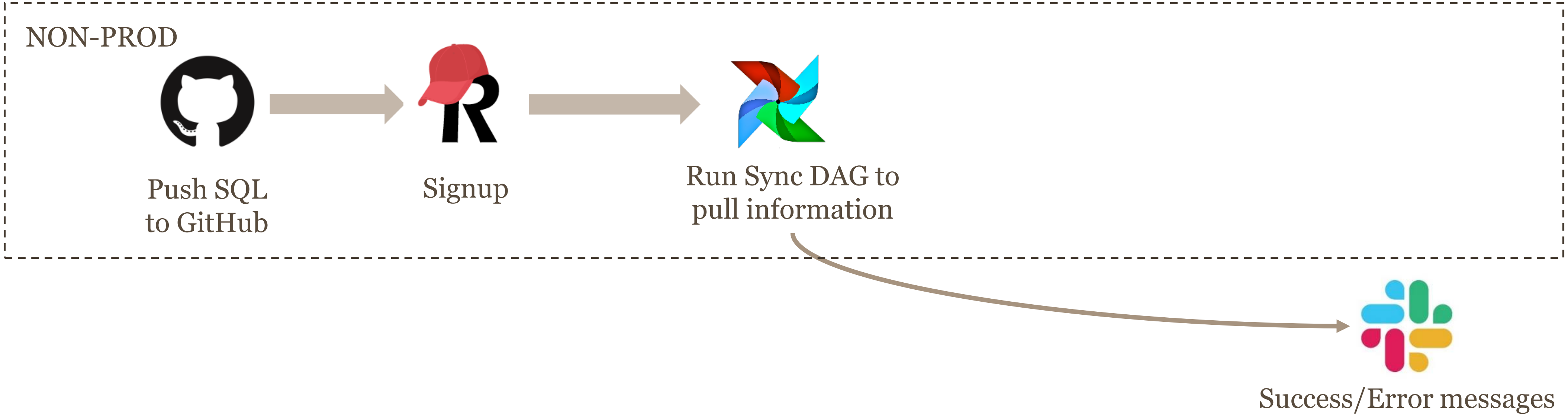
○ Existing Data Mart
○ Flat File
○ Other
○ None

Mention any external dependencies here e.g. other Fact Tables or Flat Files

CDW Dependency [DEPRECATED]

Default Dependency is CDW-Daily2. Please confirm with root pod if unsure.

AUTOMARTS • INITIAL DESIGN



NON-PROD

Push SQL to GitHub

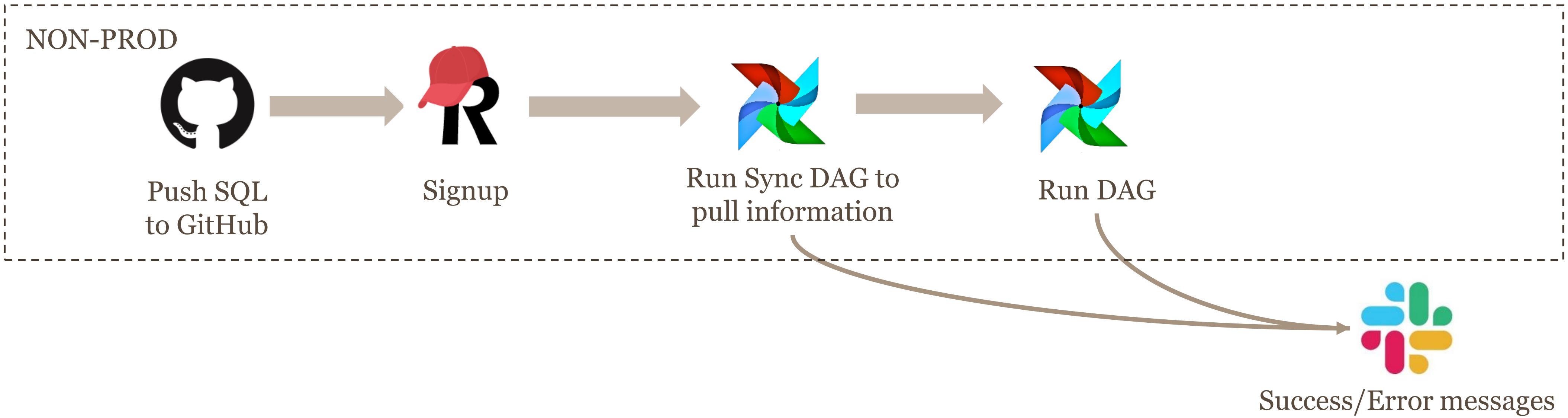
Signup

Run Sync DAG to pull information

Success/Error messages

Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

AUTOMARTS • INITIAL DESIGN



NON-PROD

Push SQL to GitHub

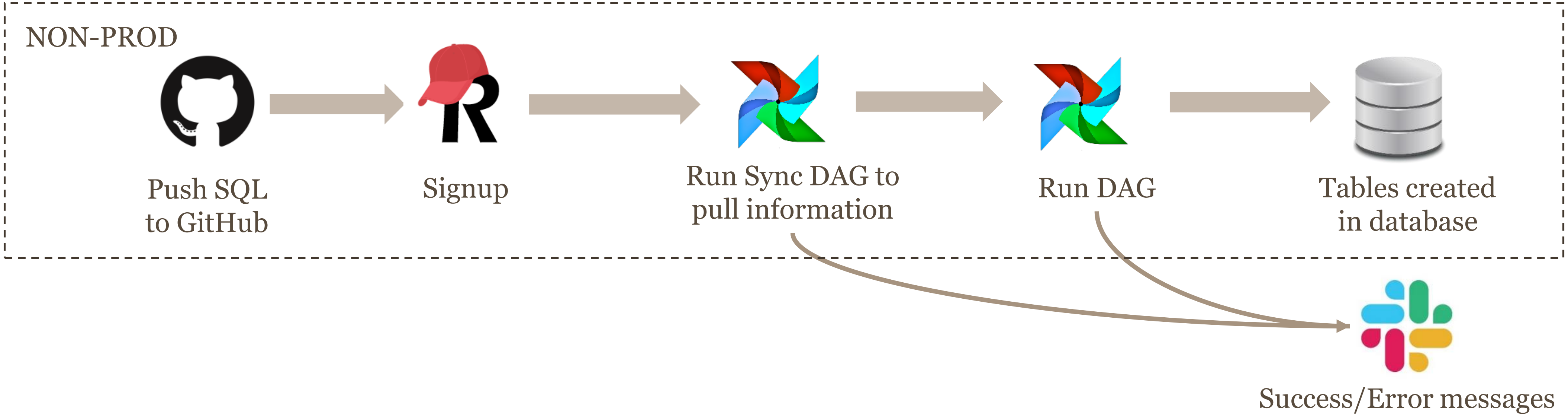
Signup

Run Sync DAG to pull information

Run DAG

Success/Error messages

AUTOMARTS • INITIAL DESIGN



NON-PROD

Push SQL to GitHub

Signup

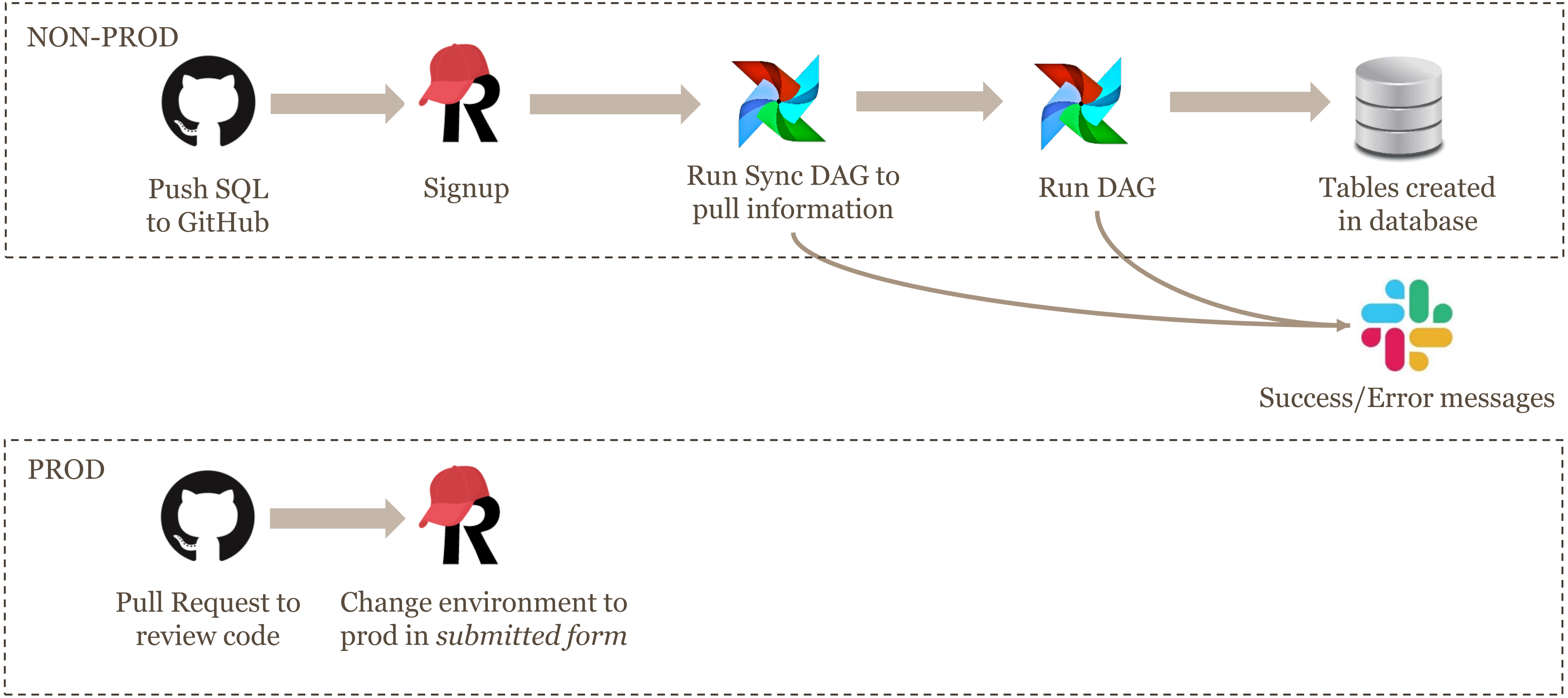
Run Sync DAG to pull information

Run DAG

Tables created in database

Success/Error messages

AUTOMARTS • INITIAL DESIGN



NON-PROD

Push SQL
to GitHub

Signup

Run Sync DAG to
pull information

Run DAG

Tables created
in database

Success/Error messages

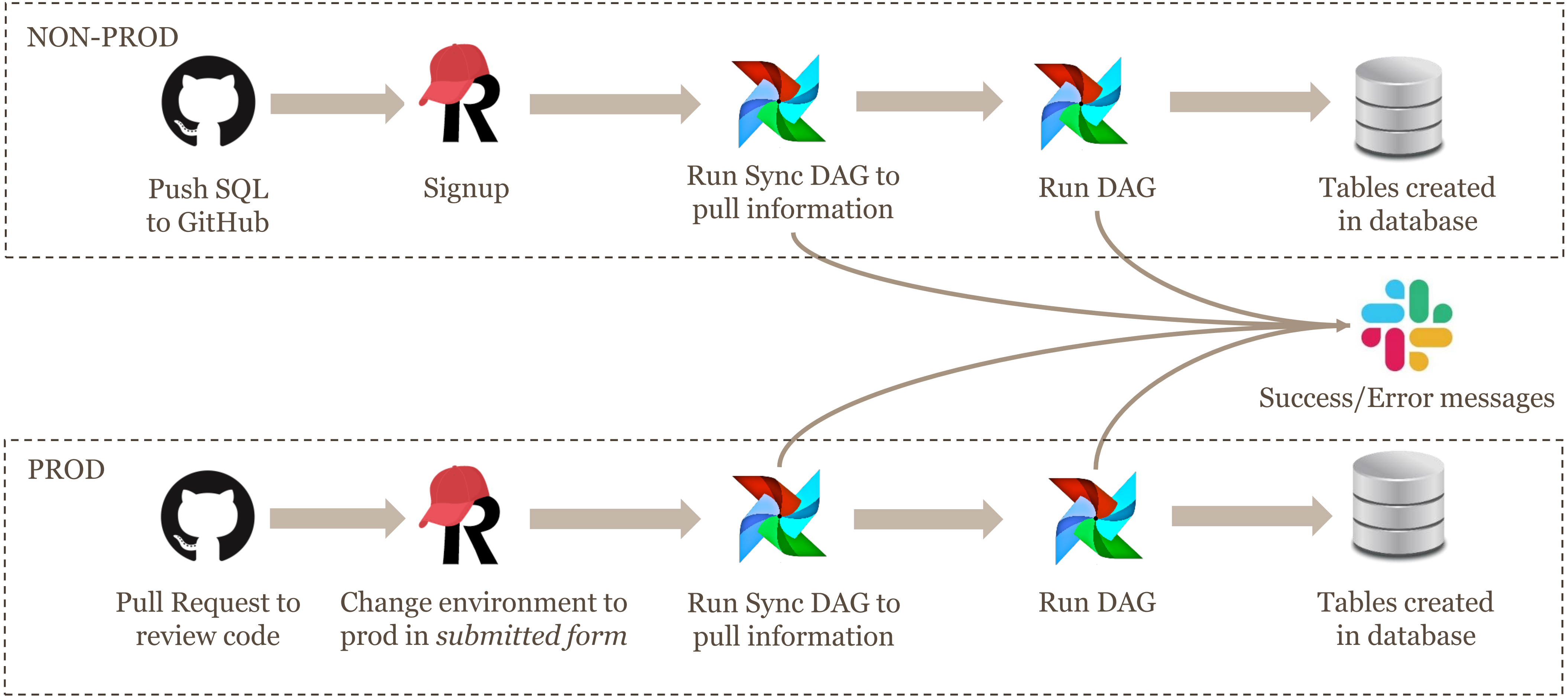
PROD

Pull Request to
review code

Change environment to
prod in *submitted form*

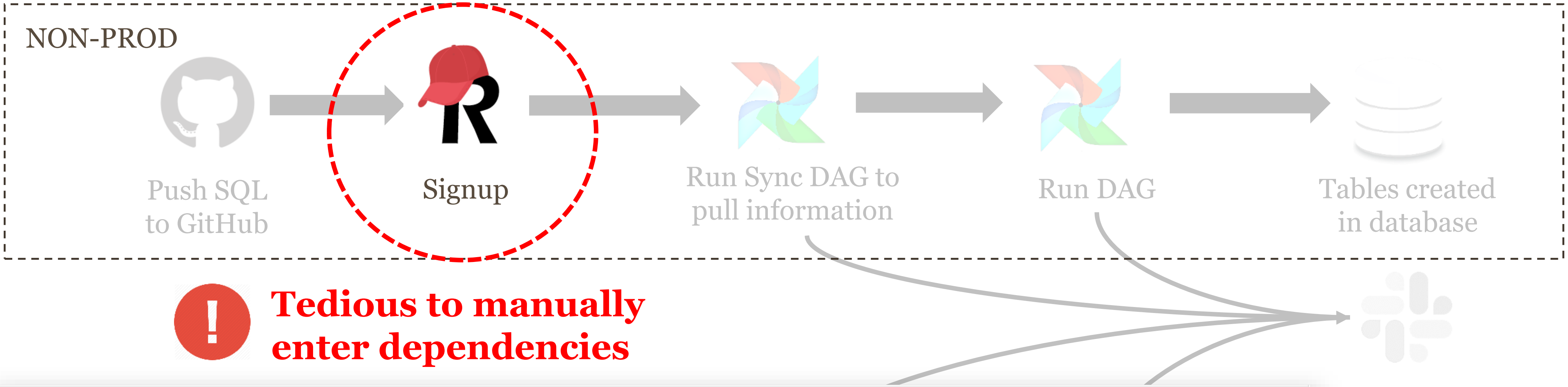
Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

AUTOMARTS • INITIAL DESIGN

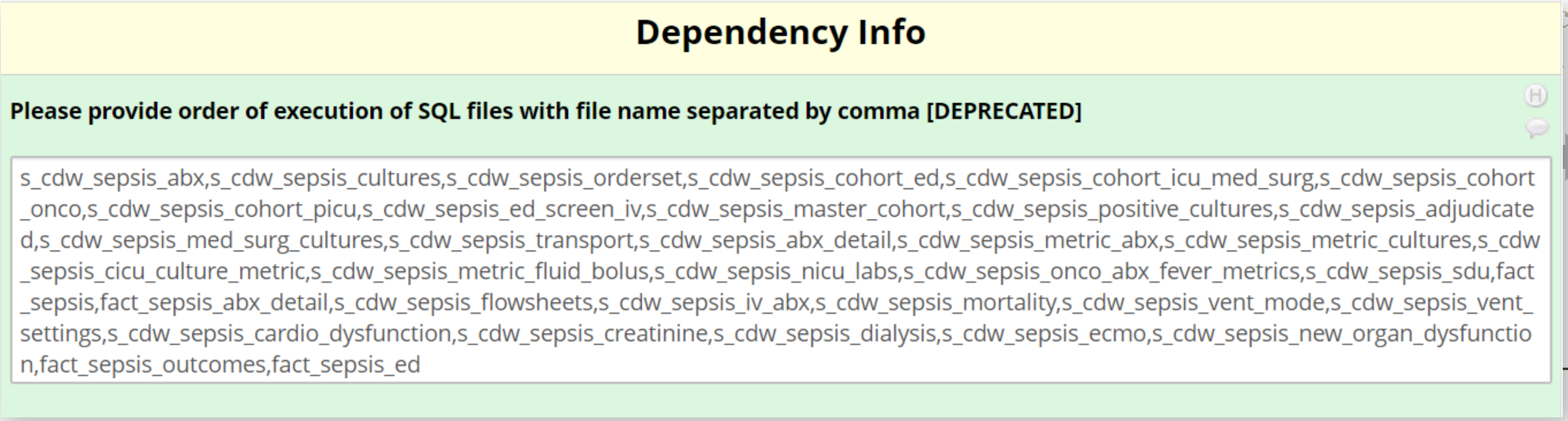

NON-PROD
Push SQL to GitHub
Signup
Run Sync DAG to pull information
Run DAG
Tables created in database
Success/Error messages
PROD
Pull Request to review code
Change environment to prod in *submitted form*
Run Sync DAG to pull information
Run DAG
Tables created in database

Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

AUTOMARTS • ISSUES WITH INITIAL DESIGN



Tedious to manually enter dependencies



Dependency Info

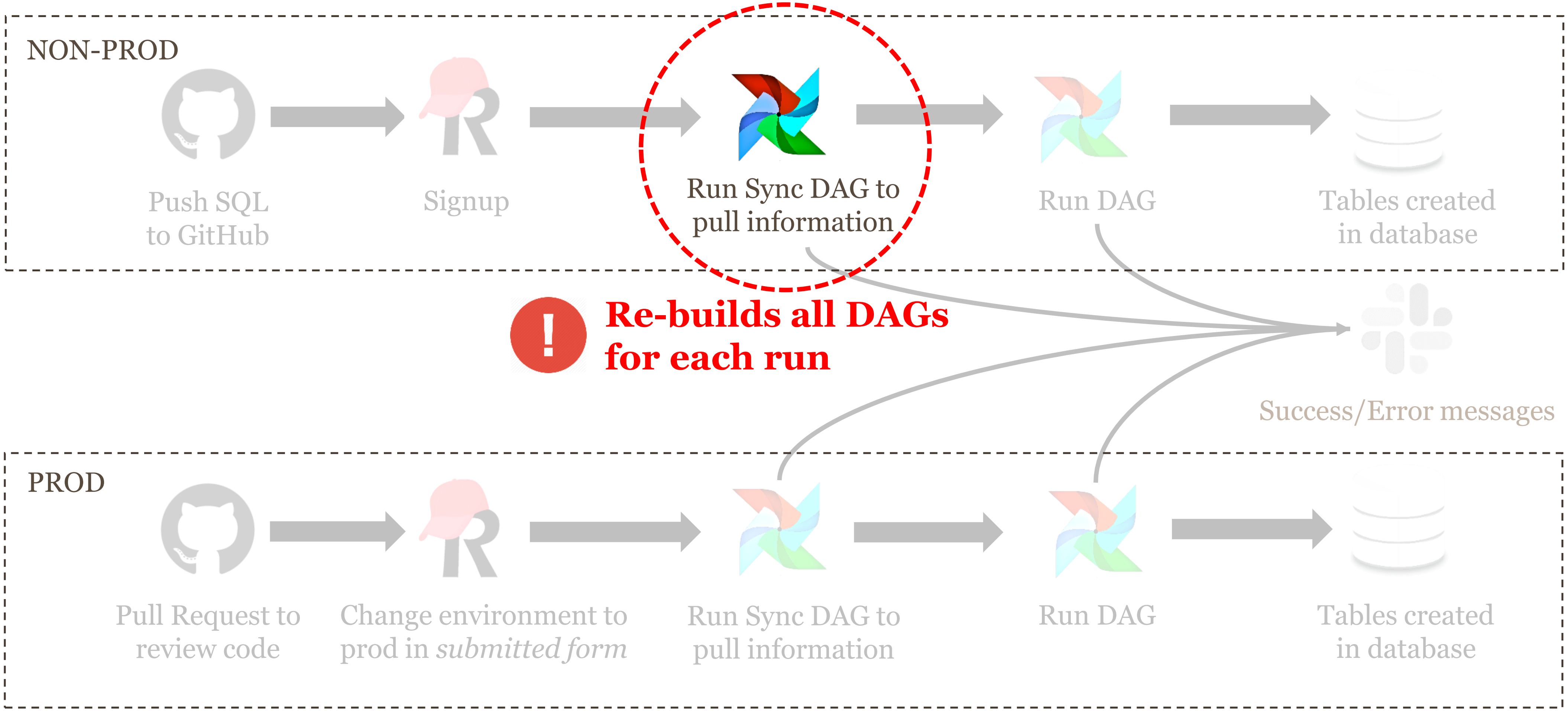
Please provide order of execution of SQL files with file name separated by comma [DEPRECATED]

```
s_cdw_sepsis_abx,s_cdw_sepsis_cultures,s_cdw_sepsis_orderset,s_cdw_sepsis_cohort_ed,s_cdw_sepsis_cohort_icu_med_surg,s_cdw_sepsis_cohort_onco,s_cdw_sepsis_cohort_picu,s_cdw_sepsis_ed_screen_iv,s_cdw_sepsis_master_cohort,s_cdw_sepsis_positive_cultures,s_cdw_sepsis_adjudicated,s_cdw_sepsis_med_surg_cultures,s_cdw_sepsis_transport,s_cdw_sepsis_abx_detail,s_cdw_sepsis_metric_abx,s_cdw_sepsis_metric_cultures,s_cdw_sepsis_cicu_culture_metric,s_cdw_sepsis_metric_fluid_bolus,s_cdw_sepsis_nicu_labs,s_cdw_sepsis_onco_abx_fever_metrics,s_cdw_sepsis_sdu,fact_sepsis,fact_sepsis_abx_detail,s_cdw_sepsis_flowsheets,s_cdw_sepsis_iv_abx,s_cdw_sepsis_mortality,s_cdw_sepsis_vent_mode,s_cdw_sepsis_vent_settings,s_cdw_sepsis_cardio_dysfunction,s_cdw_sepsis_creatinine,s_cdw_sepsis_dialysis,s_cdw_sepsis_ecmo,s_cdw_sepsis_new_organ_dysfunction,fact_sepsis_outcomes,fact_sepsis_ed
```

ess/Error messages

Tables created in database

AUTOMARTS • ISSUES WITH INITIAL DESIGN

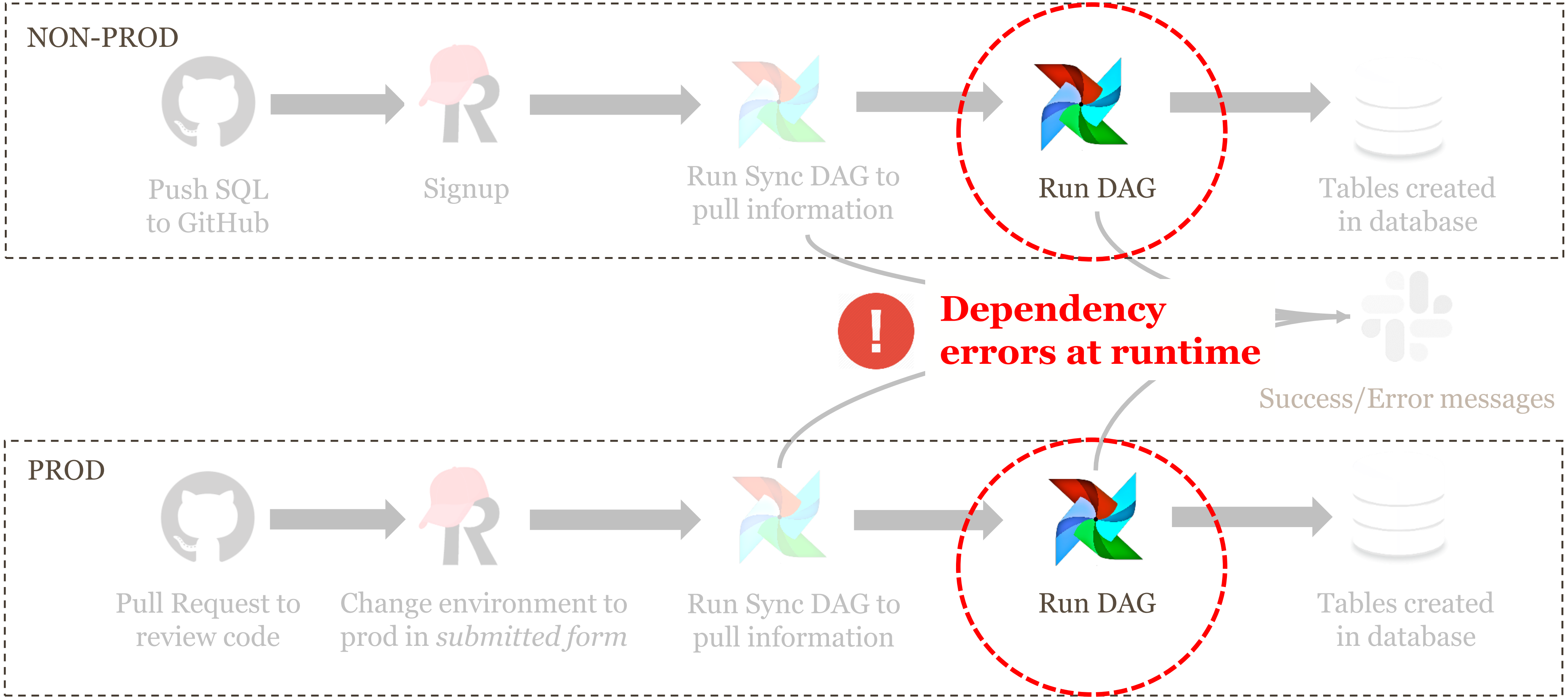


NON-PROD
Push SQL to GitHub
Signup
Run Sync DAG to pull information
Run DAG
Tables created in database
Success/Error messages
PROD
Pull Request to review code
Change environment to prod in *submitted form*
Run Sync DAG to pull information
Run DAG
Tables created in database

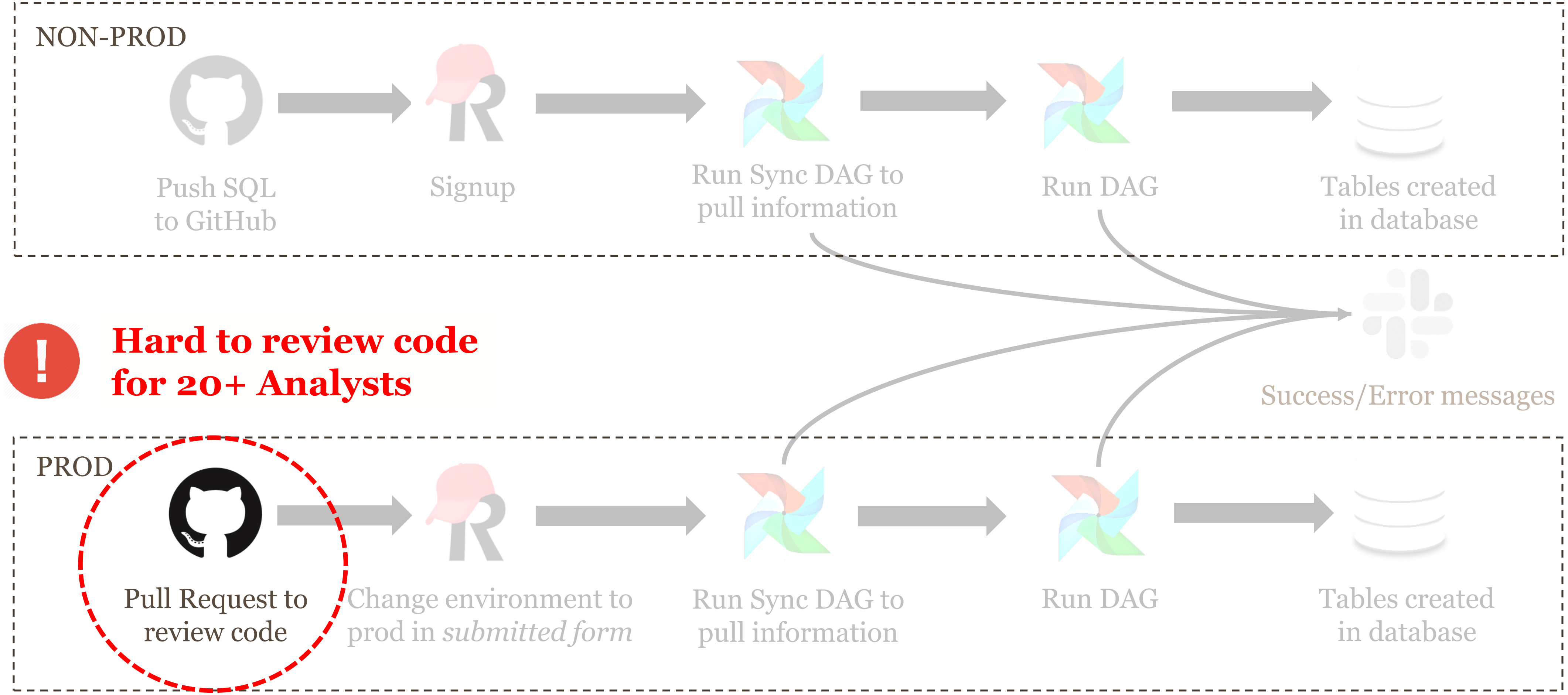
Re-builds all DAGs for each run

Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

AUTOMARTS • ISSUES WITH INITIAL DESIGN

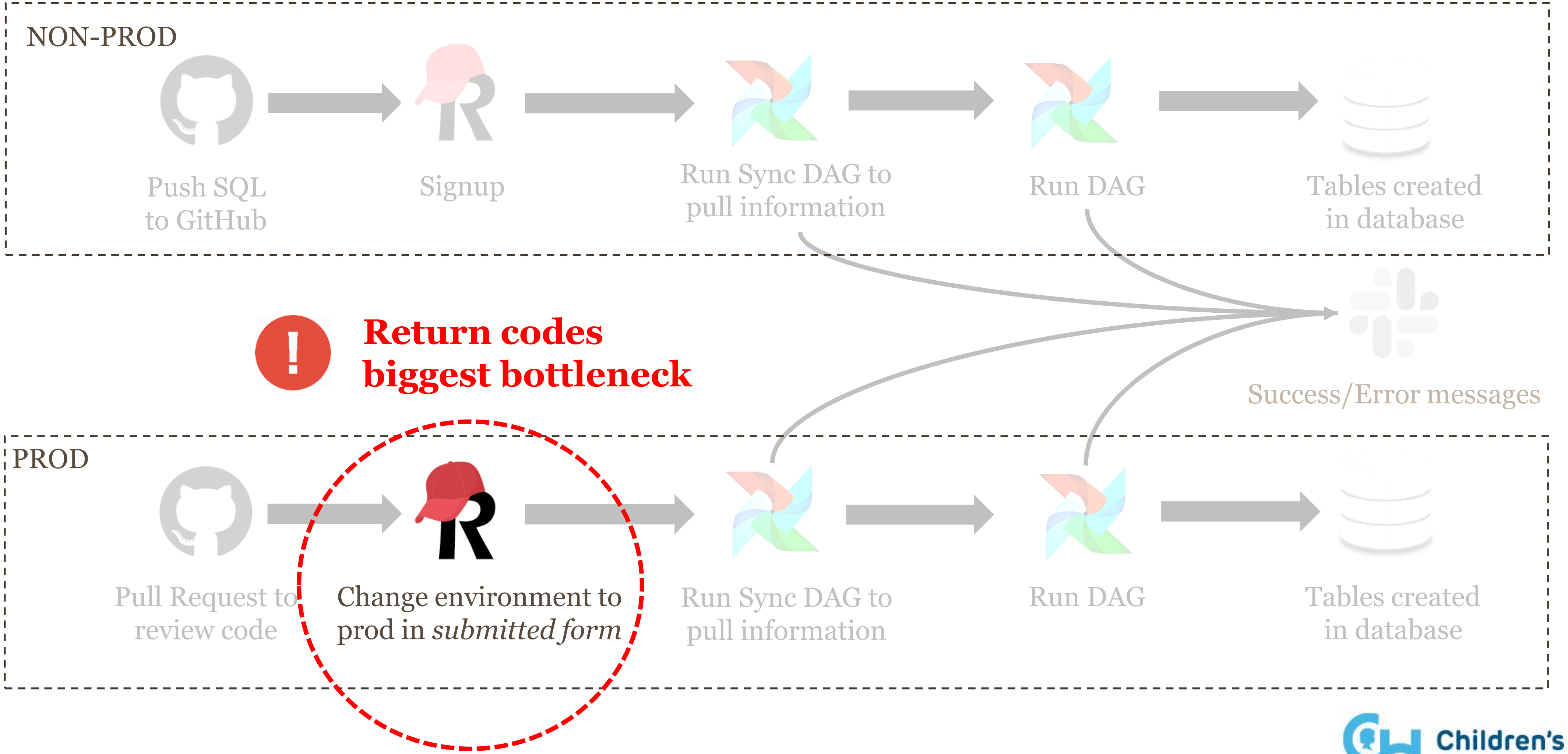

NON-PROD
Push SQL to GitHub
Signup
Run Sync DAG to pull information
Run DAG
Tables created in database
!
Dependency errors at runtime
Success/Error messages
PROD
Pull Request to review code
Change environment to prod in *submitted form*
Run Sync DAG to pull information
Run DAG
Tables created in database

AUTOMARTS • ISSUES WITH INITIAL DESIGN


NON-PROD
Push SQL to GitHub
Signup
Run Sync DAG to pull information
Run DAG
Tables created in database
Success/Error messages
PROD
Pull Request to review code
Change environment to prod in *submitted form*
Run Sync DAG to pull information
Run DAG
Tables created in database

! **Hard to review code for 20+ Analysts**

AUTOMARTS • ISSUES WITH INITIAL DESIGN



NON-PROD

Push SQL to GitHub → Signup → Run Sync DAG to pull information → Run DAG → Tables created in database

Success/Error messages

Return codes biggest bottleneck

PROD

Pull Request to review code → Change environment to prod in *submitted form* → Run Sync DAG to pull information → Run DAG → Tables created in database

Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

AUTOMARTS • OTHER ISSUES

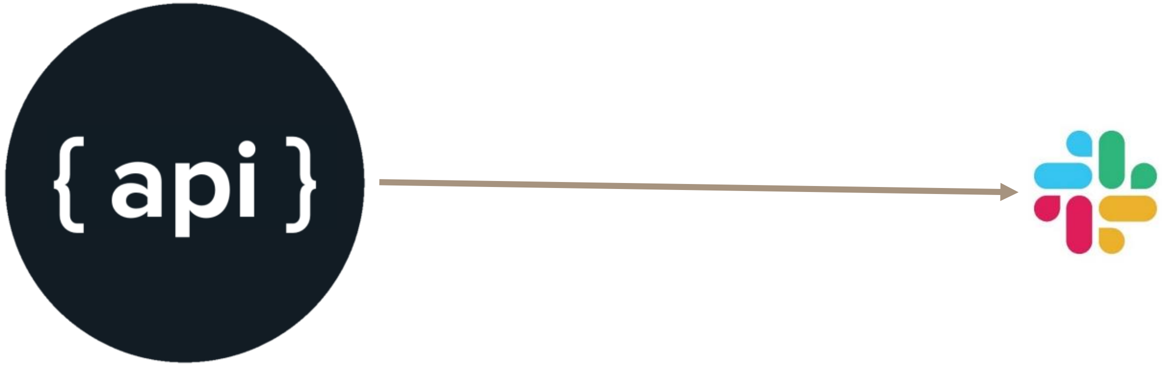
- **Scalability**
  The adoption rate was much better than we anticipated.

- **Database Performance Issues**
  No prioritization/batching of workflows resulted in 30+ SQL tasks running at a given instant

- **Airflow Webpage Slowness**
  Since all workflows kick-off at same time, webpage reload slows down significantly

AUTOMARTS • CURRENT STATE

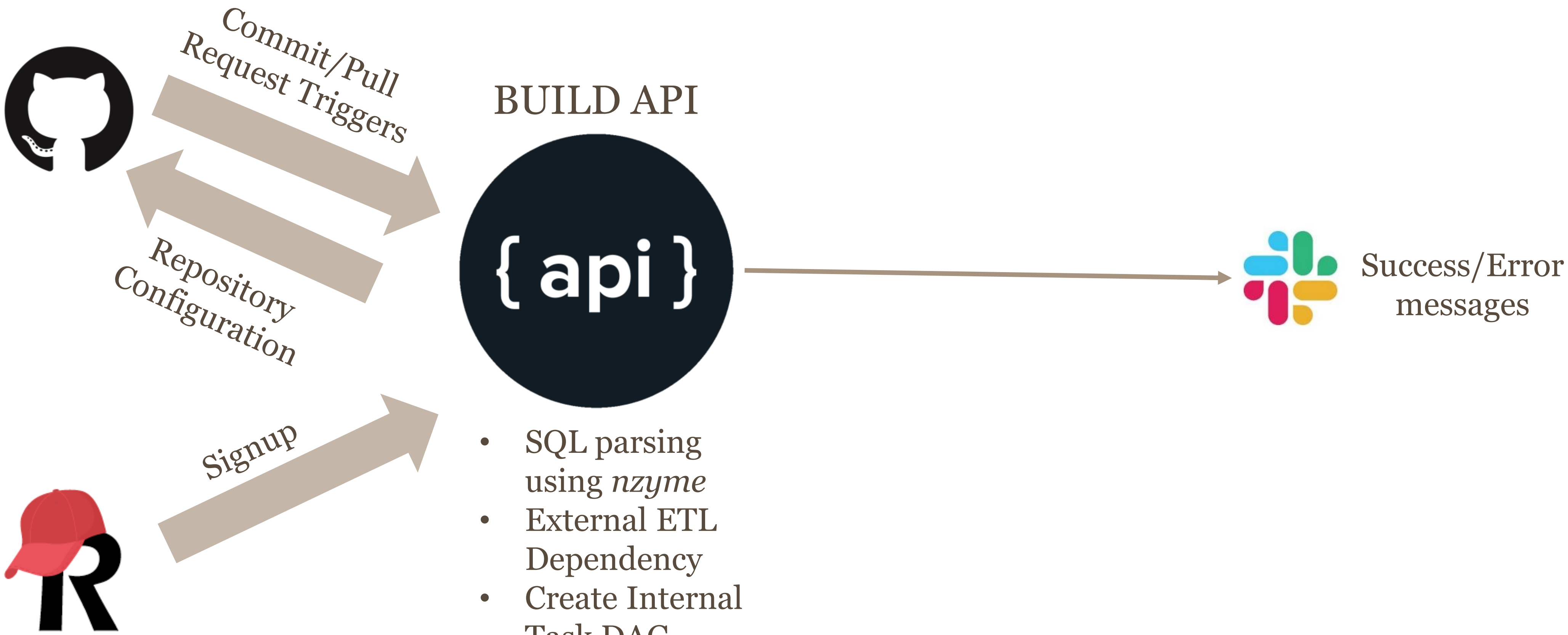
BUILD API



Success/Error messages

- SQL parsing using *nzyme*
- External ETL Dependency
- Create Internal Task DAG
- Cache SQL on server

AUTOMARTS • CURRENT STATE



Commit/Pull Request Triggers

Repository Configuration

Signup

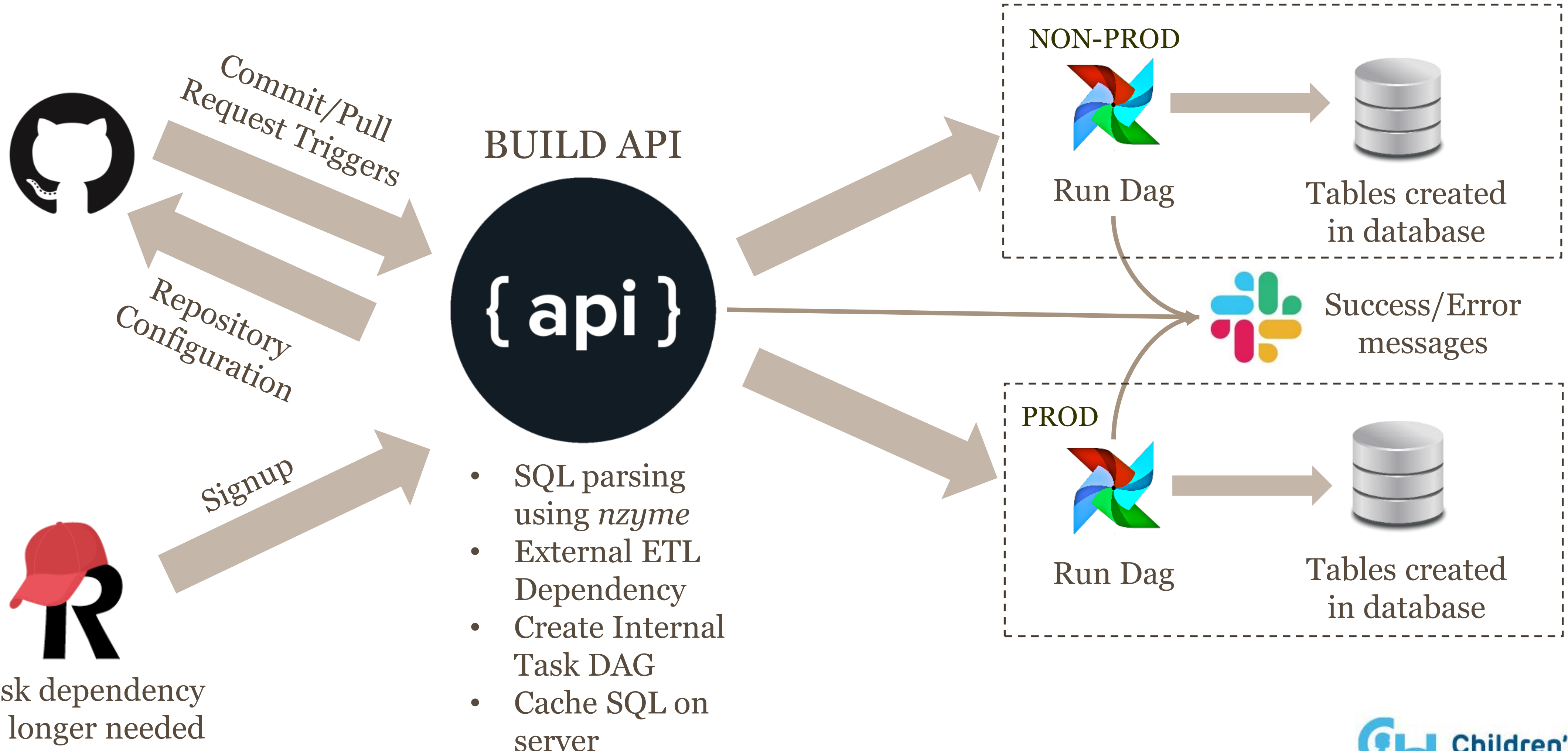
BUILD API

Success/Error messages

Task dependency no longer needed

- SQL parsing using *nzyme*
- External ETL Dependency
- Create Internal Task DAG
- Cache SQL on server

AUTOMARTS • CURRENT STATE

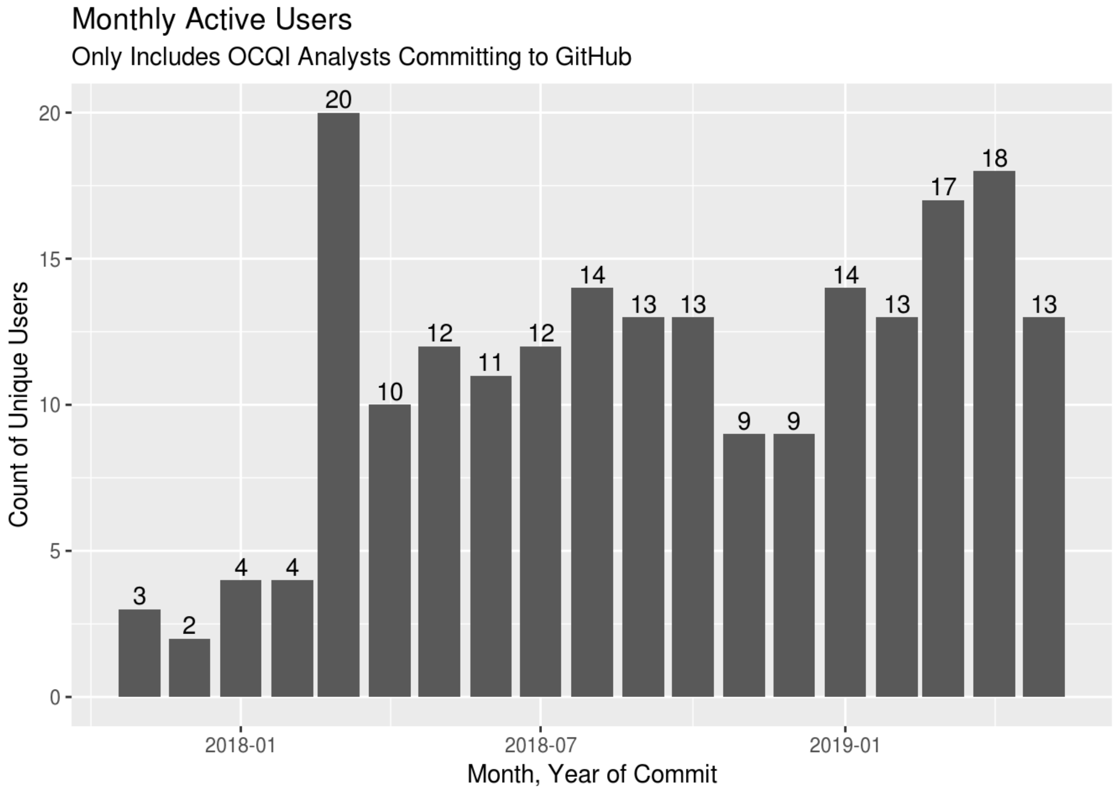

Commit/Pull Request Triggers
Repository Configuration
Signup
Task dependency no longer needed
BUILD API
{ api }
NON-PROD
Run Dag
Tables created in database
Success/Error messages
PROD
Run Dag
Tables created in database

- SQL parsing using *nzyme*
- External ETL Dependency
- Create Internal Task DAG
- Cache SQL on server

Children's Hospital of Philadelphia
Center for Healthcare Quality & Analytics

AUTOMARTS • IMPACT

- Turn around time for business logic change is minimal
- Data Engineers are able to focus on other challenging problems
- 130+ workflows in Airflow
- ~6 workflows per Analyst
- 1000+ hours time savings overall


Monthly Active Users

Only Includes OCQI Analysts Committing to GitHub

Children's Hospital of Philadelphia®
Center for Healthcare Quality & Analytics

QUESTIONS?