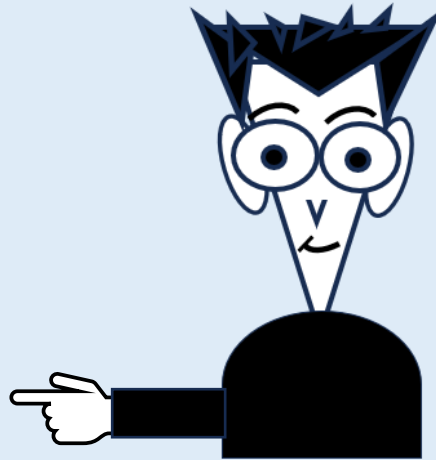


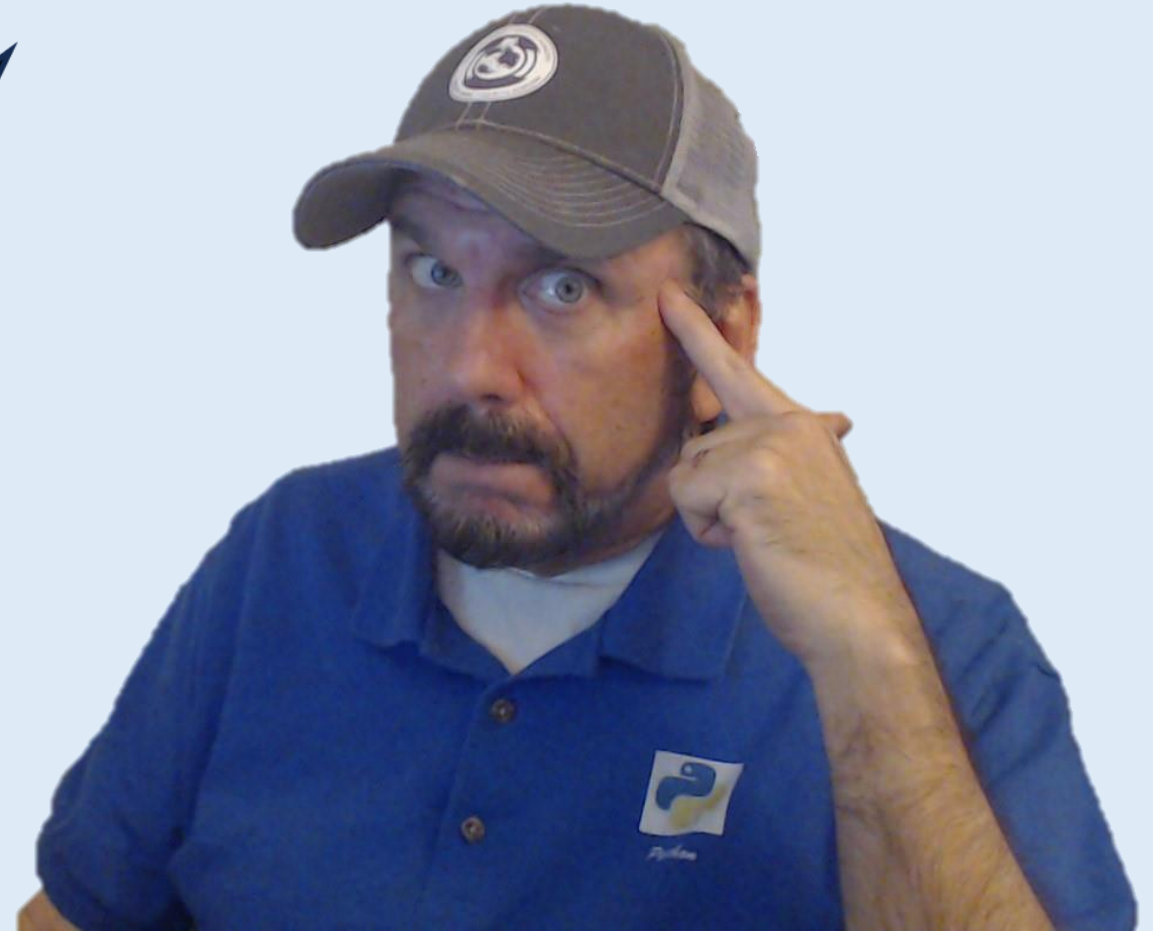
# Data Orchestrators vs. Schedulers

*Automating your Data Workloads*

Video #1



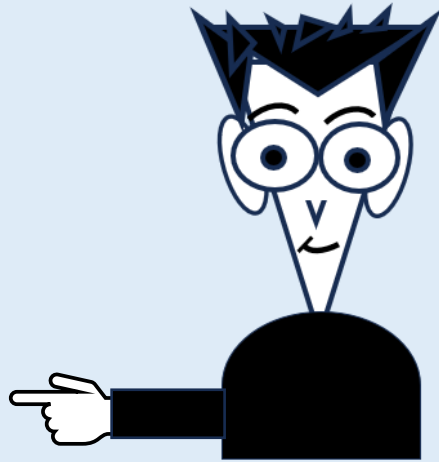
Bryan Cafferky  
YouTube Channel



# Automating Your Data Workloads

## *Schedulers vs. Orchestrators*

Video #1



# Where We're Going?

- What's the Problem?
- Job Schedulers vs. Data Orchestrators
- What is a Job Scheduler?
- What is a Data Orchestrator?
- When to Use a Job Scheduler
- When to Use a Data Orchestrator
- Wrap Up

# What's the Problem?

- **It's Becoming Cumbersome to Manually Run Some Programs**
- **You Want to Automate the Execution of the Work**
  - **ETL Job – Load Some Files into SQL Tables**
  - **Run Some Utilities Like a Spark Optimize or Repartition**
  - **Execute Reports and Send them to Users**
  - **Send a Letter to Aunt Tilly**
  - **Run Any Programs You Want**

# Two Types of Job Schedulers

- **General Service like Cron or Windows Task Scheduler**
  - Also, IBM Mainframe Scheduler, JES/Job Control Language
- **Data Platform Specific Schedulers**
  - SQL Server Agent
  - Databricks Workflows
  - Azure Data Factory

# What is a Job Scheduler?

- Service that Automates the Execution of a Job which consists of a Task or Set of Tasks.
- When the Job Will Be Executed is Configurable.
- Common Job Triggers:
  - *Time Schedule like daily, weekly, hourly.*
  - *Based on an Event like a file landing in a folder.*
  - *Jobs Sometimes May Call Other Jobs*

# Generic Schedulers

# Example: Cron

```
root@utls-newsletter-php7:~ # crontab -l
# do daily/weekly/monthly maintenance
# min    hour    day    month    weekday  command
*/15     *      *      *      *        run-parts /etc/periodic/15min
0        *      *      *      *        run-parts /etc/periodic/hourly
0        2      *      *      *        run-parts /etc/periodic/daily
0        3      *      *      6        run-parts /etc/periodic/weekly
0        5      1      *      *        run-parts /etc/periodic/monthly

22 0 * * * "/root/.acme.sh"/acme.sh --cron --home "/root/.acme.sh" > /dev/null
*/5 * * * * curl -L -s https://newsletter.cyberciti.biz/scheduled.php > /dev/null 2>&1
root@utls-newsletter-php7:~ #
```

← cron list jobs command

Execute run-parts

Job  
Schedules

© www.cyberciti.biz

<https://www.cyberciti.biz/faq/linux-show-what-cron-jobs-are-setup/>



# Example: Windows Task Scheduler

**Edit Trigger**

Begin the task: On a schedule

**Settings**

☐ One time

☒ Daily

☐ Weekly

☐ Monthly

Start: 9/ 7/2014 12:06:00 AM ☒ Synchronize across time zones

Recur every: 1 days

**Advanced settings**

☐ Delay task for up to (random delay): 1 hour

☒ Repeat task every: 2 minutes for a duration of: 1424 minutes

☐ Stop all running tasks at end of repetition duration

☐ Stop task if it runs longer than: 3 days

☐ Expire: 9/ 7/2015 10:06:13 AM ☐ Synchronize across time zones

☒ Enabled

OK Cancel

Schedule

time interval

When to Stop

# Data Platform Specific Schedulers

# Example: SQL Server Agent

New Job

Select a page

- General
- Steps
- Schedules
- Alerts
- Notifications
- Targets

Connection

Server:

Connection:

[View connection properties](#)

Progress

Ready

Script Help

Job step list:

Step	Name	Type	On Success	On Failure
1	Daily Status Report	Task	Continue the job reporting success	Quit the job reporting failure
2	Weekend Update Report	Task	Continue the job reporting success	Quit the job reporting failure
3	Sunday Weekly Transaction Re...	Task	Continue the job reporting success	Quit the job reporting failure

Move step:

Start step:

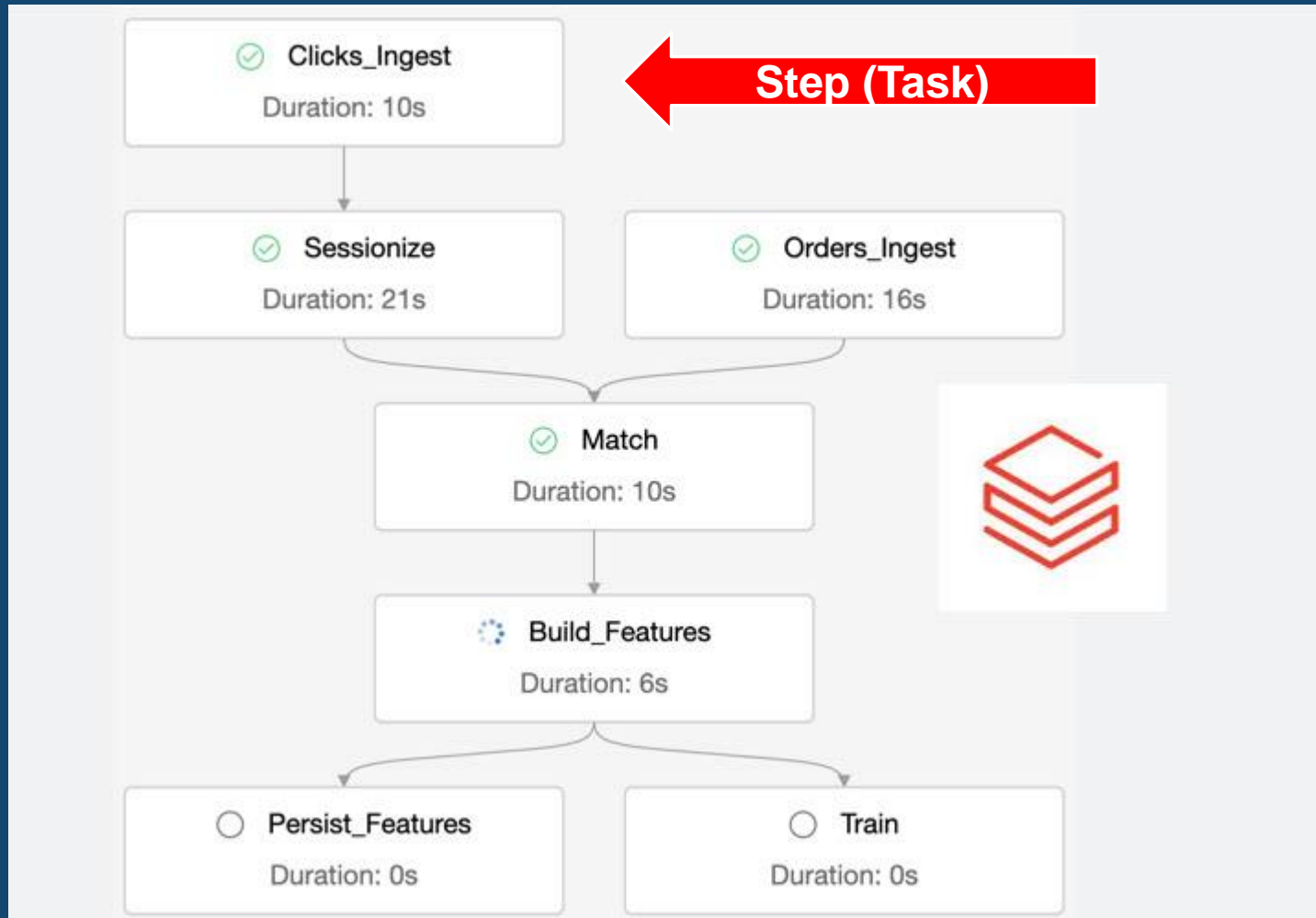
1:Daily Status Report

New... Insert... Edit Delete

OK Cancel

Job  
Definition

# Example: Databricks Workflow



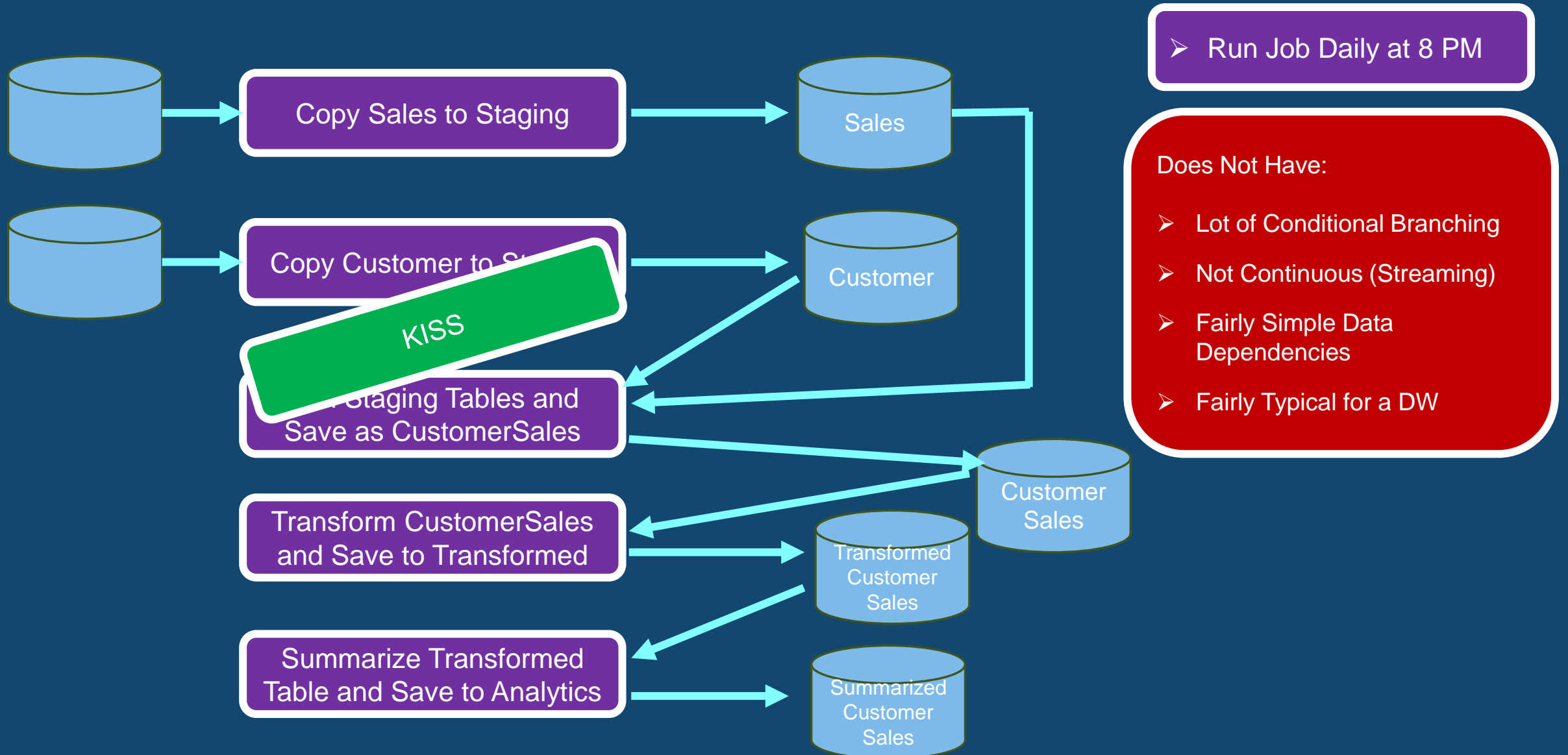
Cross Between an  
Orchestrator and a  
Scheduler

Job  
Definition

# Popular Schedulers

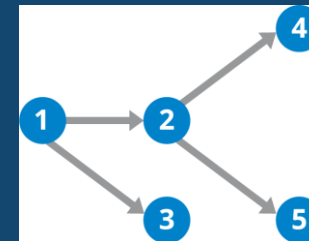
- **Cron on Linux:** Typically Executes Bash or Python scripts.
- **Windows Task Scheduler:** Executes Batch files or PowerShell scripts.
- **SQL Server Agent:** Executes SQL Services and Command/PowerShell scripts.
- **Commercial: CA7**
- **Cloud:**
  - **Azure Automation:** Executes PowerShell, Python 2 or Python 3.
  - **Azure Batch**
  - **Databricks Workflows.**
  - **Azure Functions?**

# When to Use a Scheduler?



# What is an Orchestrator?

- **Fairly New Concept! Scheduler on Steroids!**
- **Supports Complexity:**
  - Virtually Unlimited Number of Tasks
  - Tasks Dependencies
  - Task Branching
  - Many Layers and Job Paths
  - Advanced Monitoring and Notifications
  - Flexible Job Restart Options
- **Usually Uses a Directed Acyclic Graph.**



# What is a Data Orchestrator?

- **Orchestrator on Data Steroids! Data Aware!**
- **Ideally Supports:**
  - Comprehensive Data Lineage
  - Data Documentation and Reporting
  - Data Quality Integration
  - Data Metrics Evaluation for Job Stop/Continuation.
  - Advanced Triggers: Streaming?
  - Extensible to Include 3<sup>rd</sup> Part Data Services
- **Usually Requires Python Programming**



# Popular Data Orchestrators



Hybrid

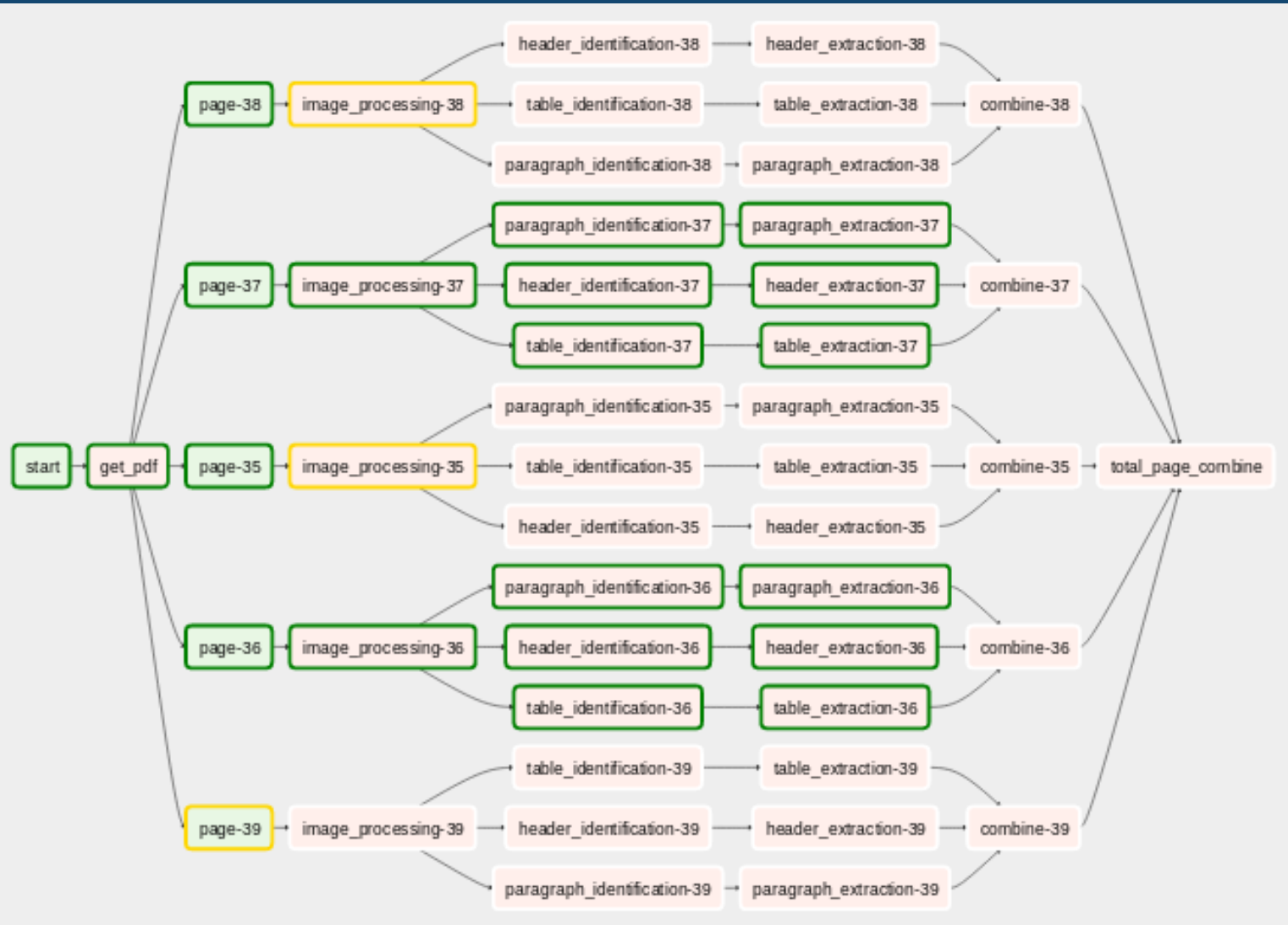
Databricks Workflows

Azure Data Factory

# When to Use a Data Orchestrator?

- **With Great Power Comes Great Responsibility!**
- **Complexity! More Work!**
- **Many Data Dependencies / Lots of Conditional Logic**
- **You Need Complete Data Transparency**
  - **Data Lineage / Reporting**
  - **Data Quality**
  - **Detailed Monitoring and Tracking**
- **You Are Willing to Commit to Python**
- **You Are Willing to Lose Some Data Platform Integration**

# When to Use a Data Orchestrator?



# Wrapping Up

- What's the Problem?
- Job Schedulers vs. Data Orchestrators
- What is a Job Scheduler?
- What is a Data Orchestrator?
- What is a Job Scheduler?
- What is a Data Orchestrator?

Thank You!