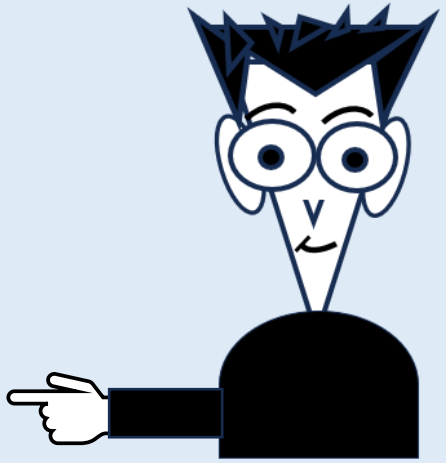
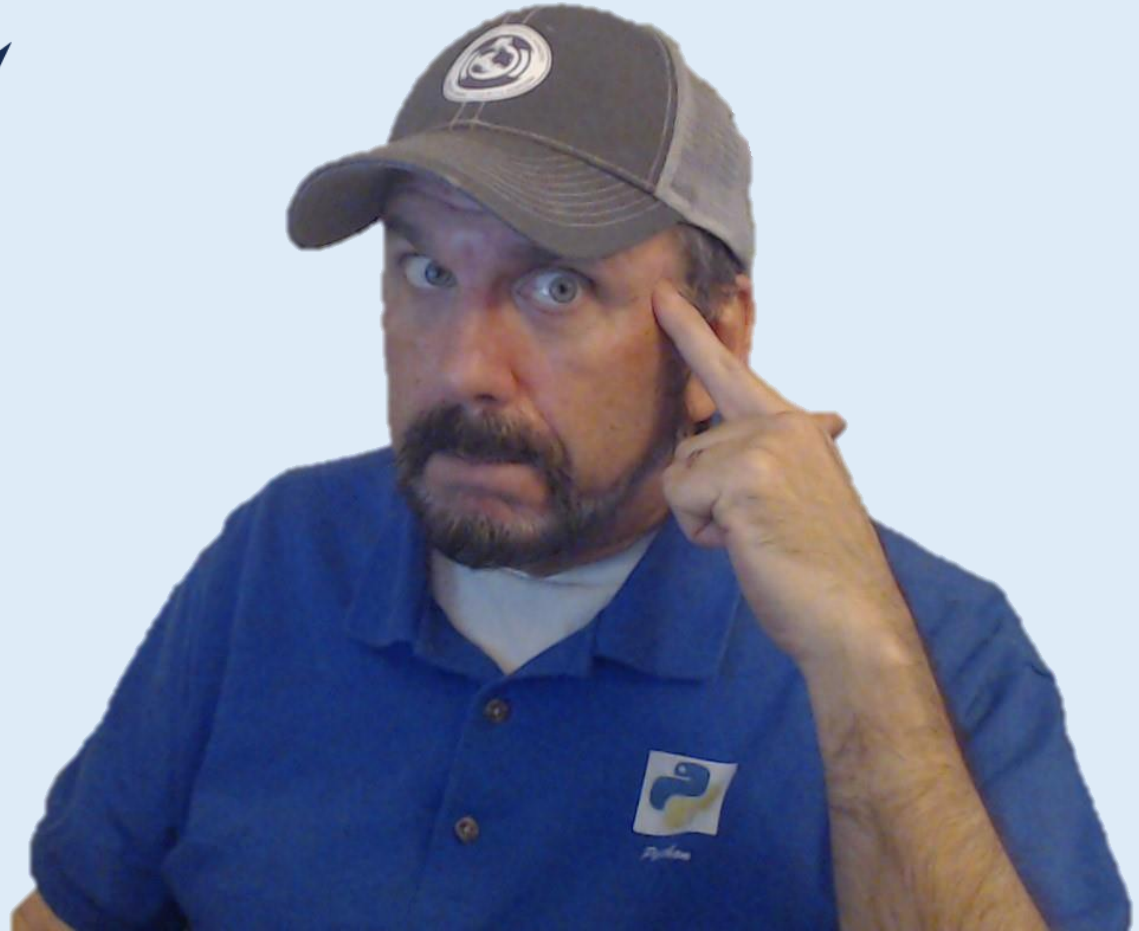
Data Orchestrators vs. Schedulers

Automating your Data Workloads

Video #1



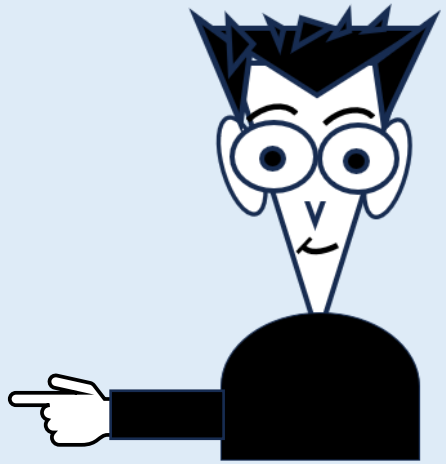


Bryan Cafferky
YouTube Channel

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

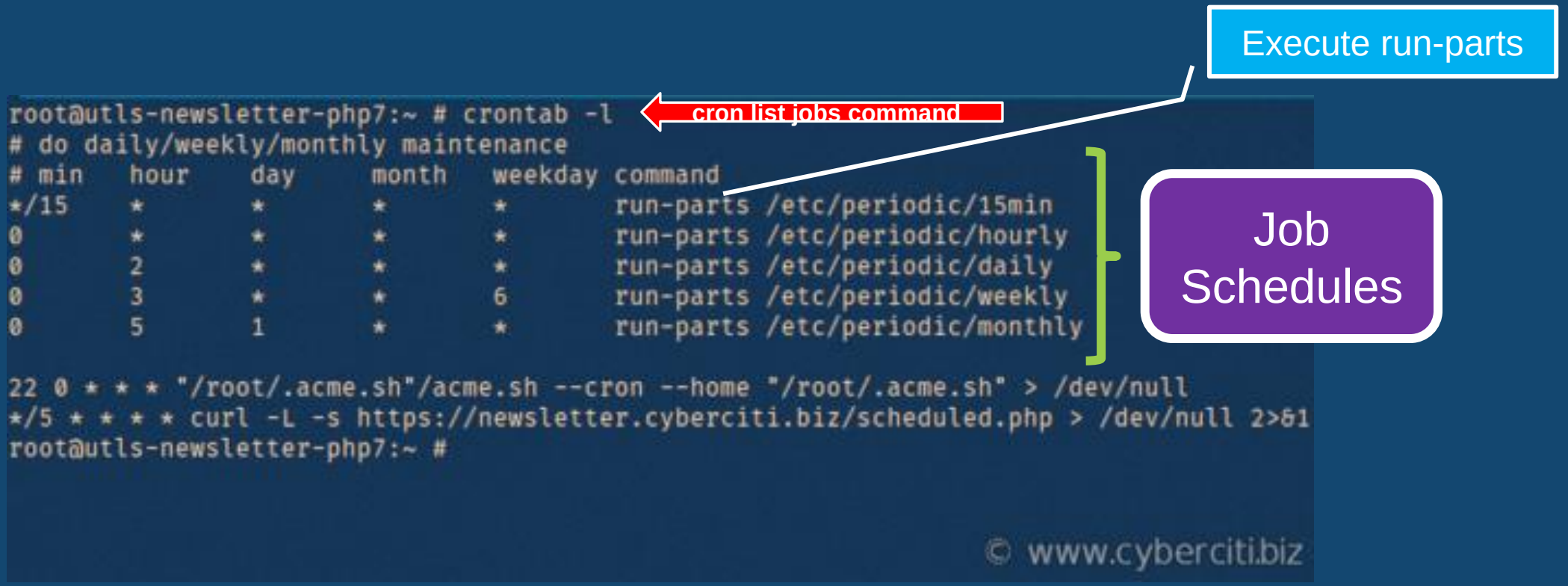
- General Service like Cron or Windows Task Scheduler
  - Also, IBM Mainframe Scheduler, JES/Job Control Language
- Data Platform Specific Schedulers
  - SQL Server Agent
  - Databricks Workflows
  - Azure Data Factory

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***

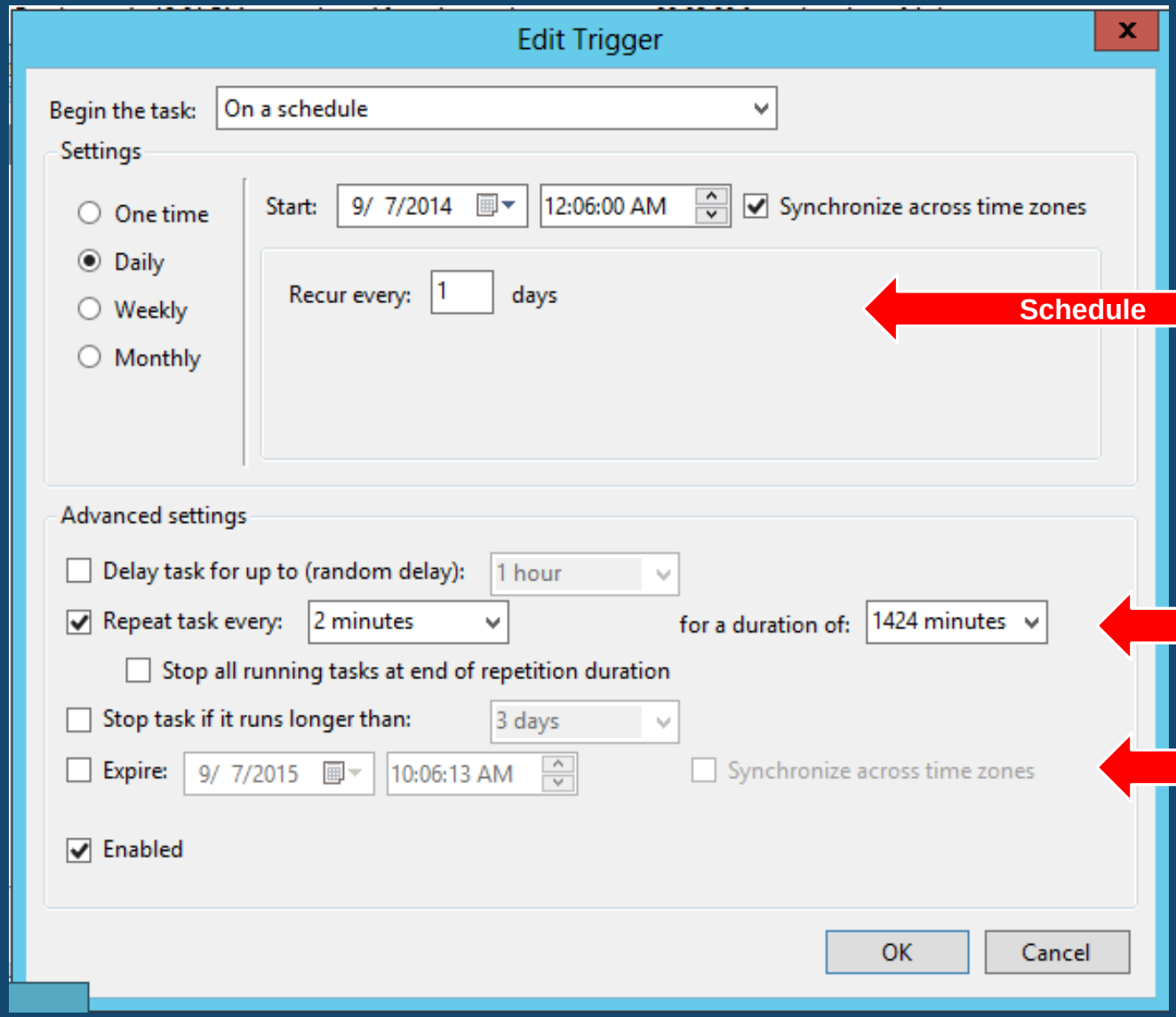
Generic Schedulers

Example: Cron



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

Example: Windows Task Scheduler



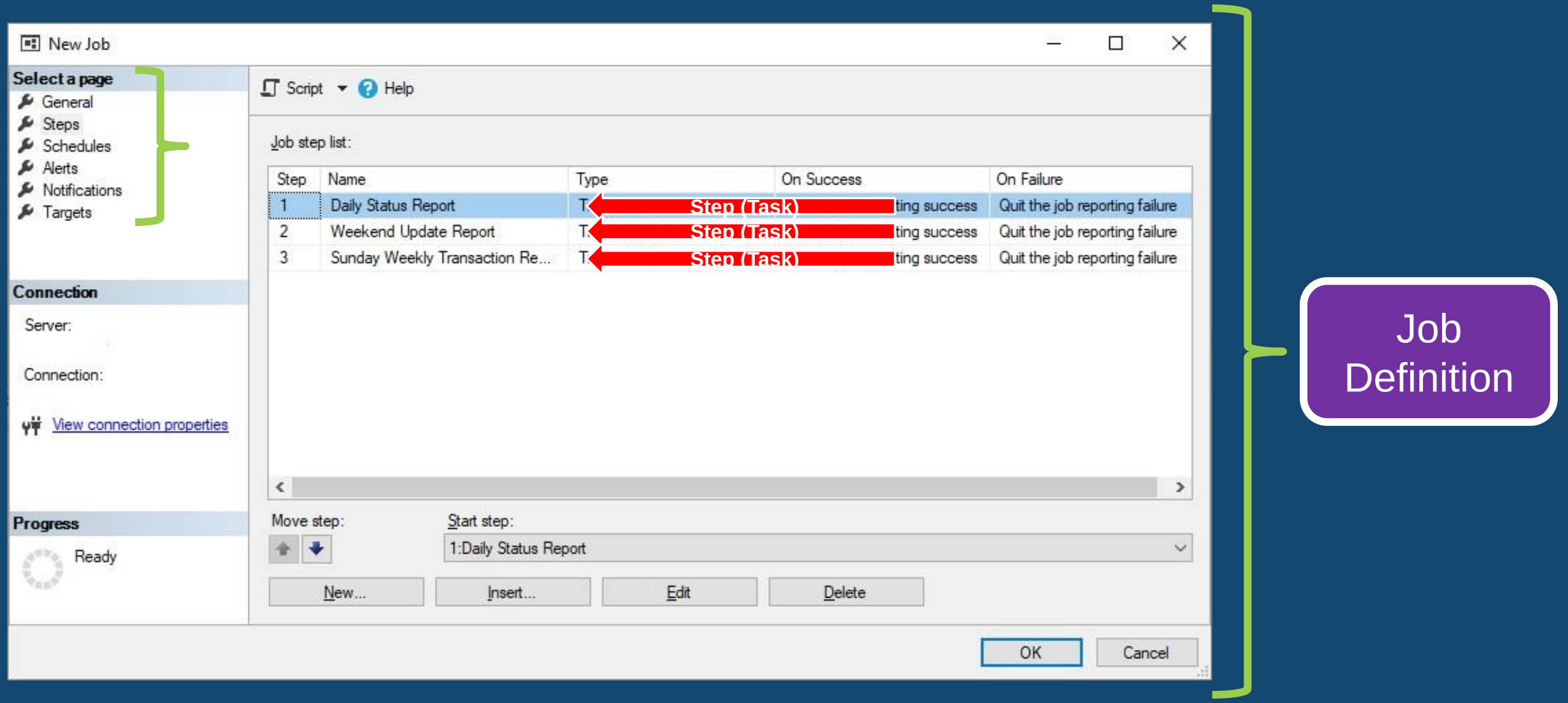
Schedule

time interval

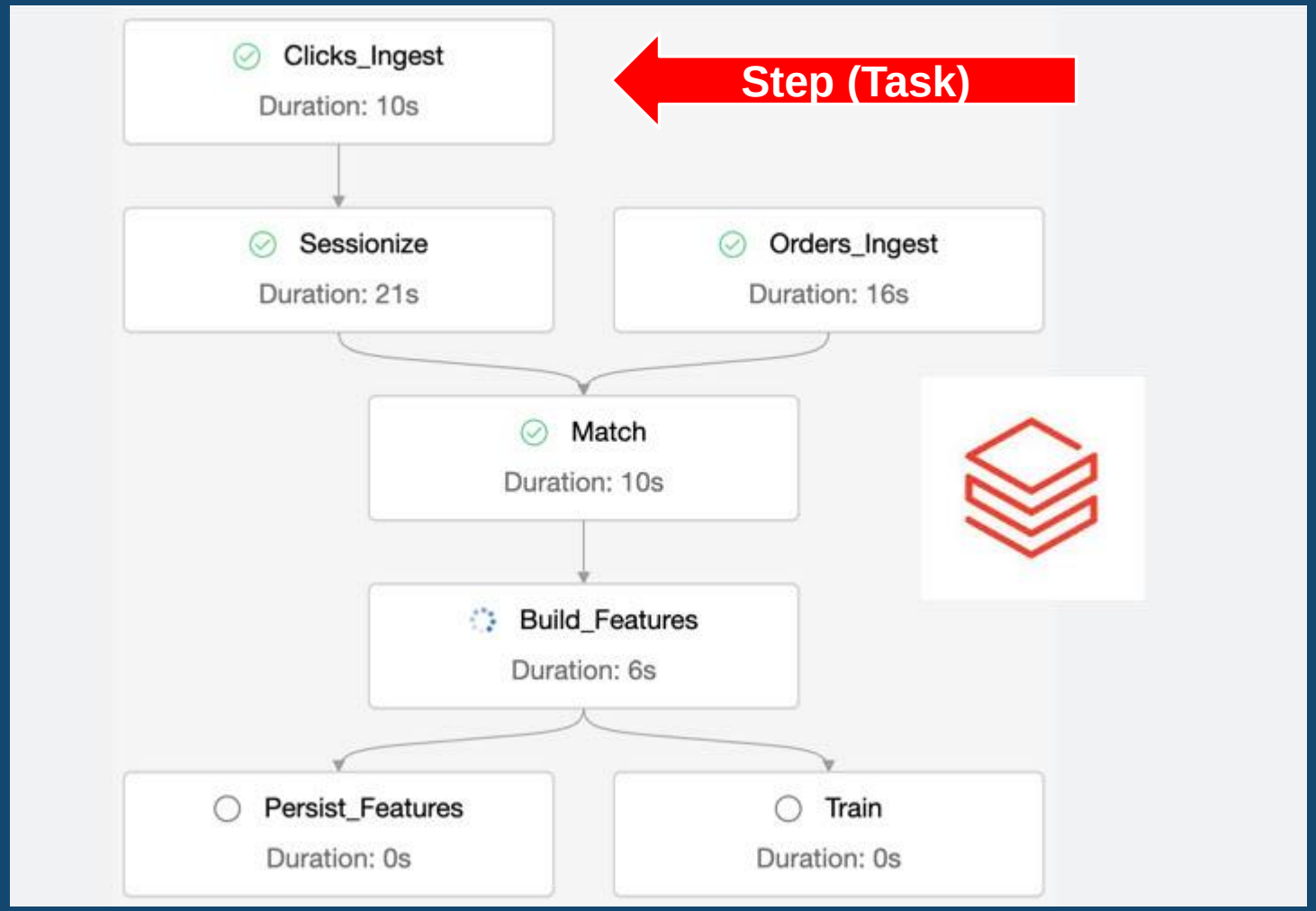
When to Stop

Data Platform Specific Schedulers

Example: SQL Server Agent



Example: Databricks Workflow



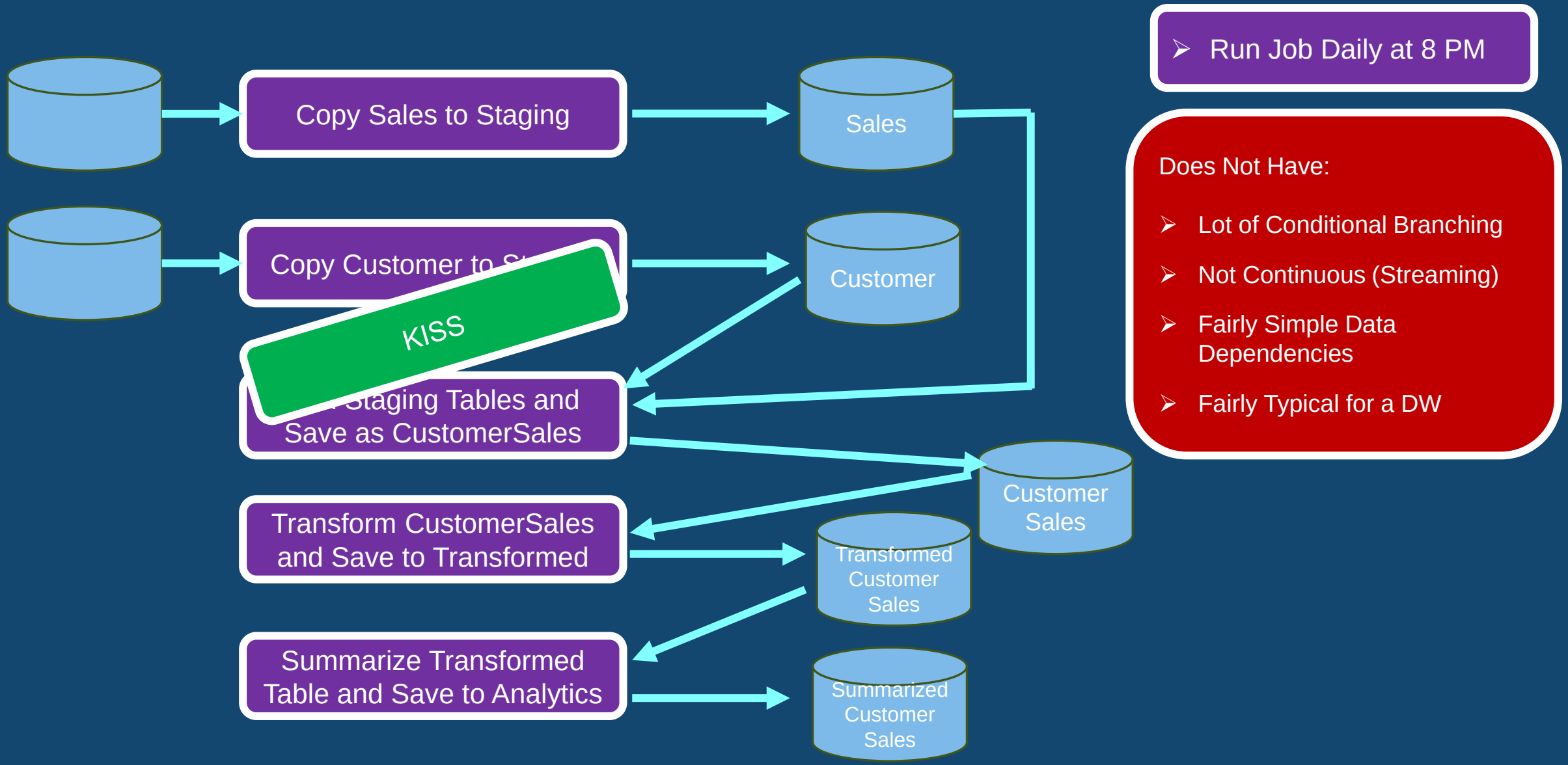
Cross Between an Orchestrator and a Scheduler

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?

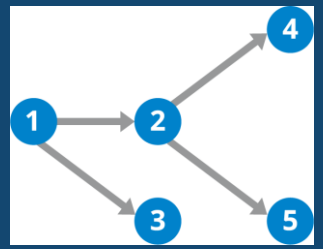


What is an Orchestrator?

- **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.**



Apache Airflow

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 3rd Part Data Services**

- **Usually Requires Python Programming**

Popular Data Orchestrators

Hybrid

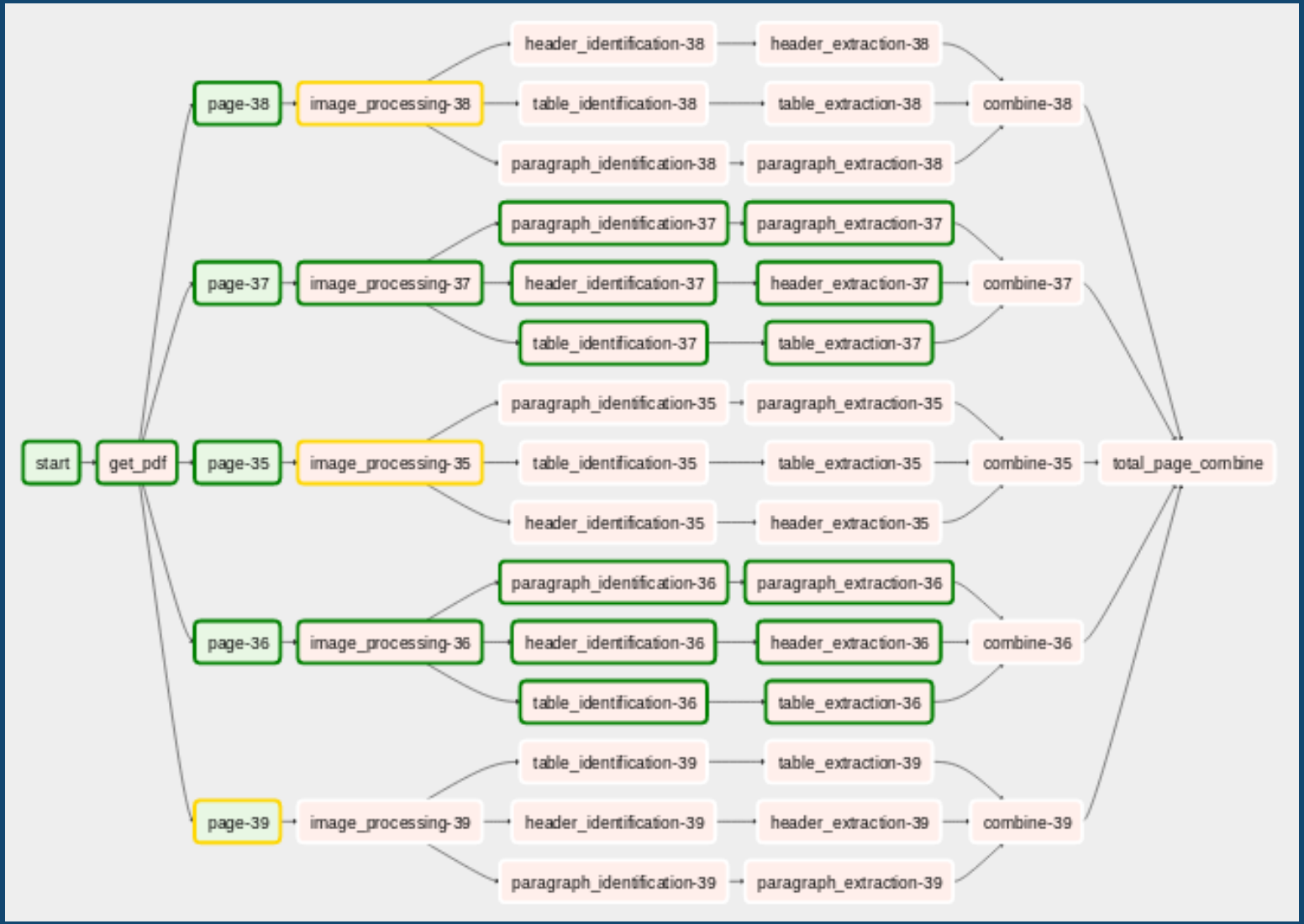
Databricks Workflows

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?
- W[illegible]cheduler
- W[illegible]Orchestrator

Thank You!