An abstract graphic featuring a series of thin, dark, wavy lines that flow from the bottom left towards the top right, creating a sense of movement and depth. The lines are layered, with some appearing in front of others, creating a three-dimensional effect. In the upper right quadrant, the word "CHHATTISGARH" is written in a bold, black, serif font, partially overlapping the wavy lines. The background is a solid, light gray.

BACHELOR OF TECHNOLOGY (HONORS)

Data Science

(COMPUTER SCIENCE & ENGINEERING)

By

Saurabh Bharti

B.Tech.(Hons) 7th Semester

Roll No: 300012821041

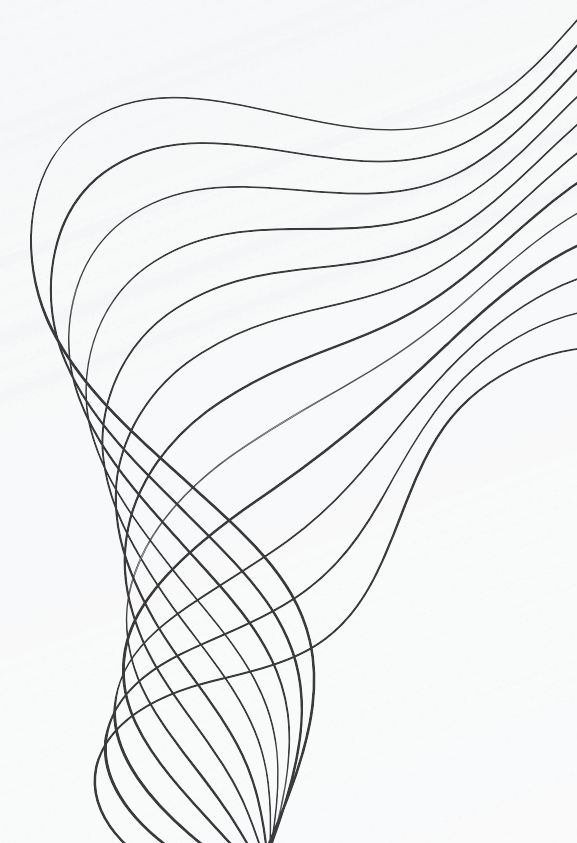
Enrollment No: CB4688

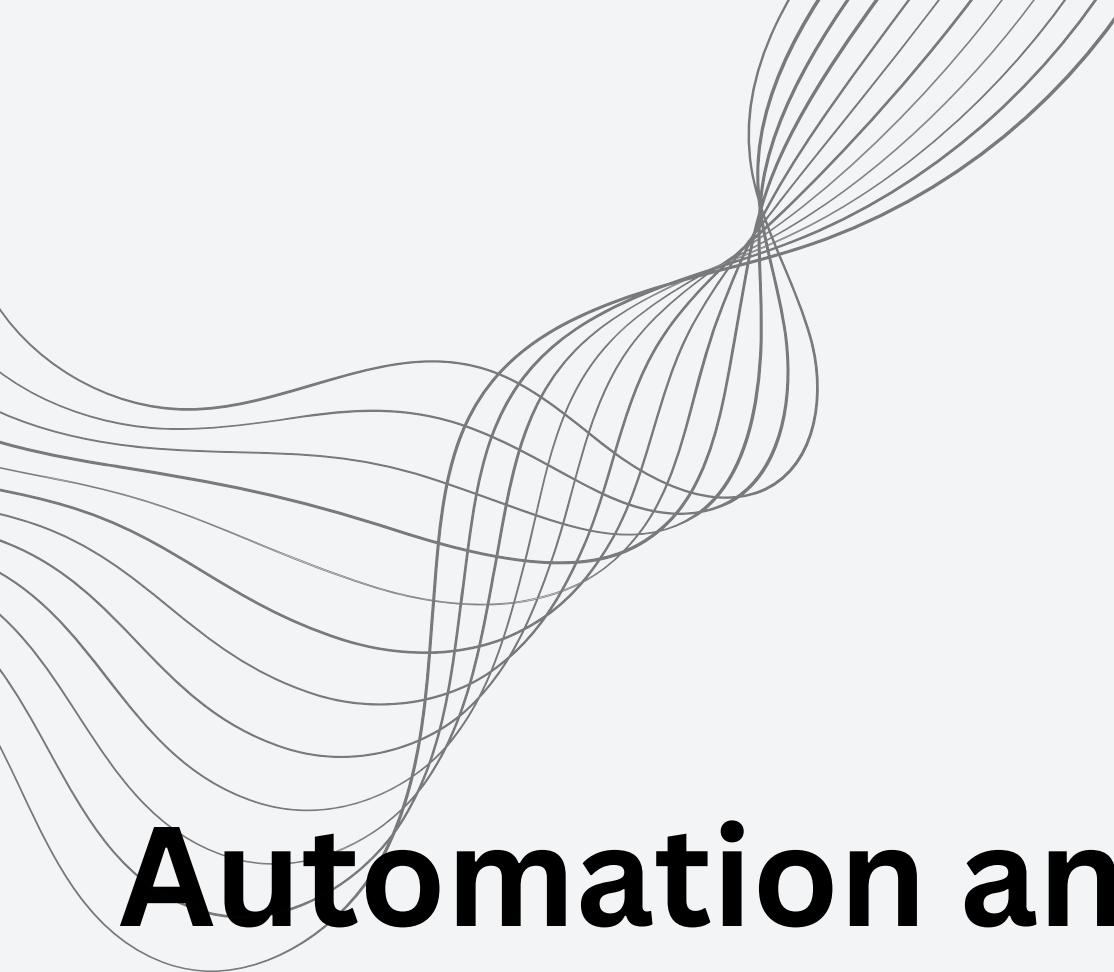
Under the Guidance of

Thaneshwari ma'am

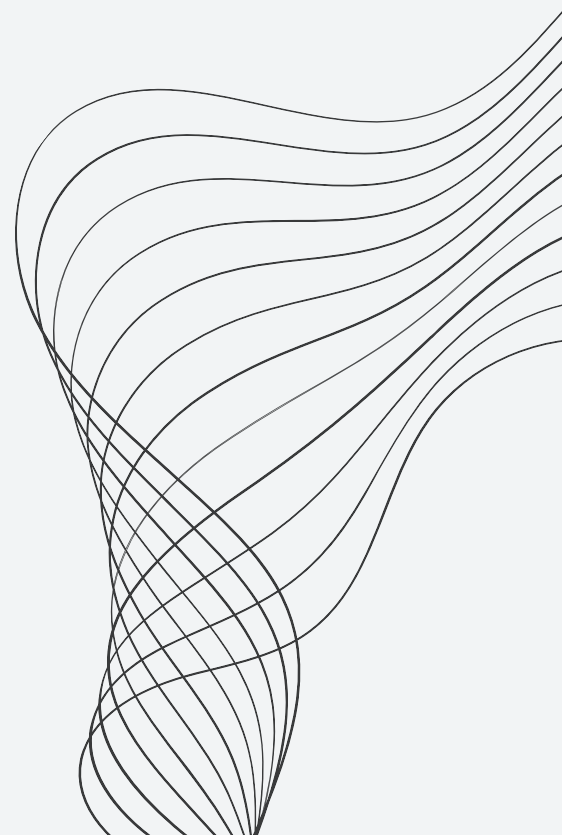
Assistant Professor (DS)

Computer Science and Engineering CSVTU, Bhilai (CG)



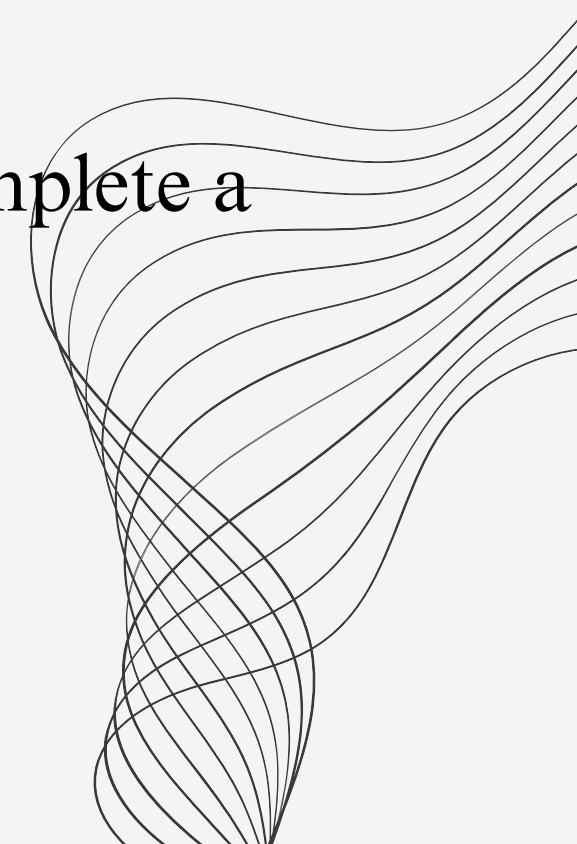


Automation and Reproducibility Using Scripting and Workflow Management Tools





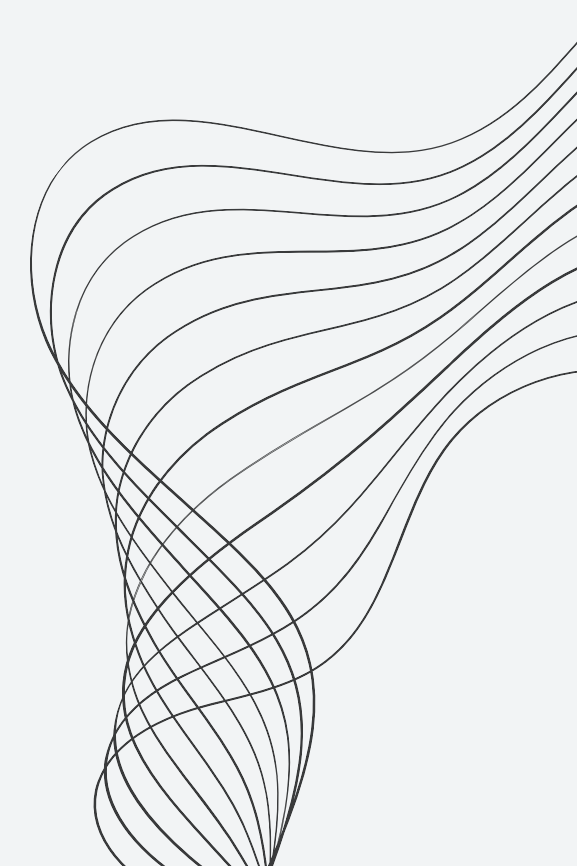
Terms in workflow automation

- **DAG (Directed Acyclic Graph):** A graph structure where nodes represent tasks, and edges represent dependencies between tasks. In workflow automation, DAGs are used to ensure tasks run in the correct sequence.
 - **Scheduler:** A component in workflow tools that runs tasks based on a schedule (like daily, weekly, or monthly).
 - **Task:** A specific job in a workflow (e.g., load data, clean data).
 - **Pipeline:** A series of connected tasks in a workflow, where each task's output is passed as input to the next task.
 - **Script** is a file containing a sequence of instructions or code that is executed
 - **Workflow** is a series of steps or tasks organized to accomplish a specific goal or complete a particular process.
- 



Example

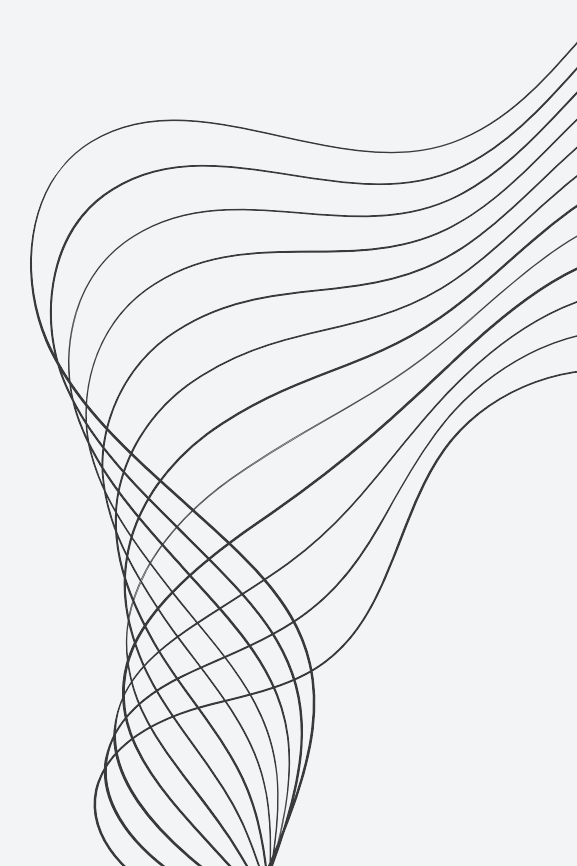
Suppose you have a company's sales data that arrives every month. This data often contains unwanted information (like manual errors, uncleaned data, etc.), and you need to process it and generate a report. Cleaning this data manually every month can be tedious. So, we'll write a script that will automatically perform the same tasks each time the new data arrives.





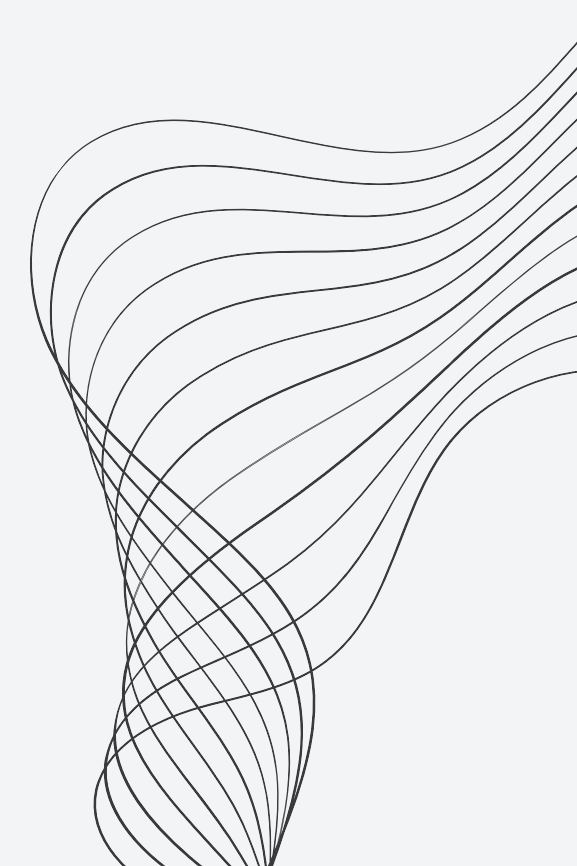
What is Automation?

Automation is the process of using technology, such as software scripts or machines, to perform tasks with minimal human intervention. It aims to streamline processes by reducing or eliminating repetitive, manual actions, allowing tasks to be completed more quickly, accurately, and consistently.





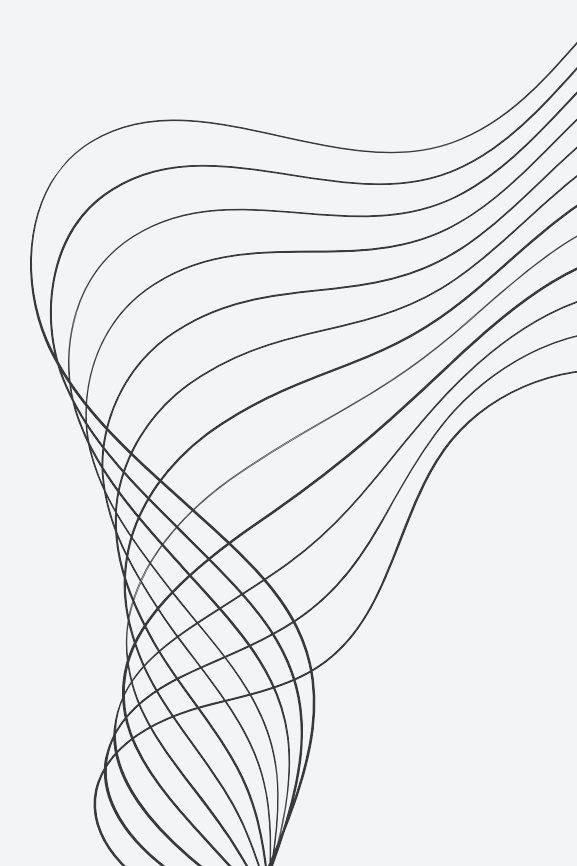
Applications

- It focuses on streamlining complex, often repetitive tasks through programming scripts and sophisticated tools that define, schedule, and monitor workflows
 - Workflows can be consistently executed to yield the same results, thus supporting both efficiency and reliability
 - Data extraction, transformation, analysis, and reporting are handled
- 



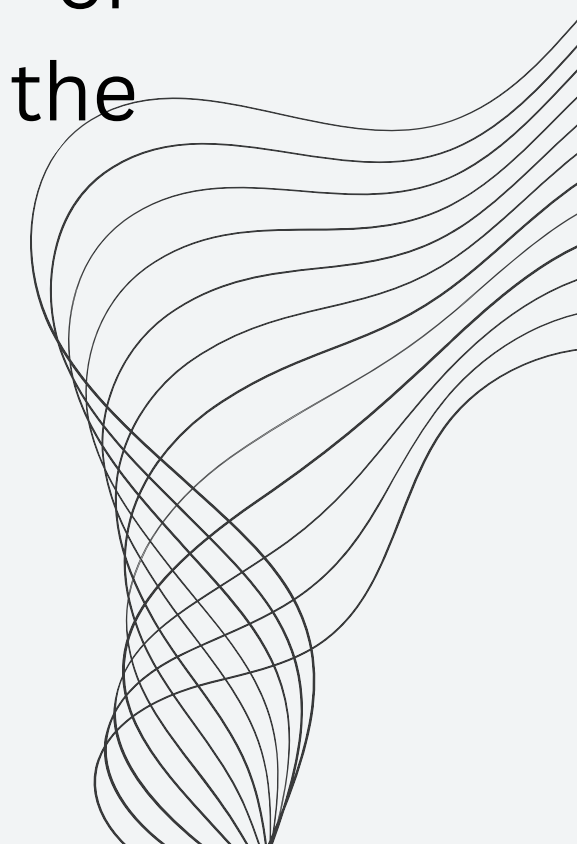
What is Reproducibility ?

Reproducibility in automated workflows refers to the ability to consistently replicate the same process and obtain identical outcomes, which is crucial in data science and analytical work.



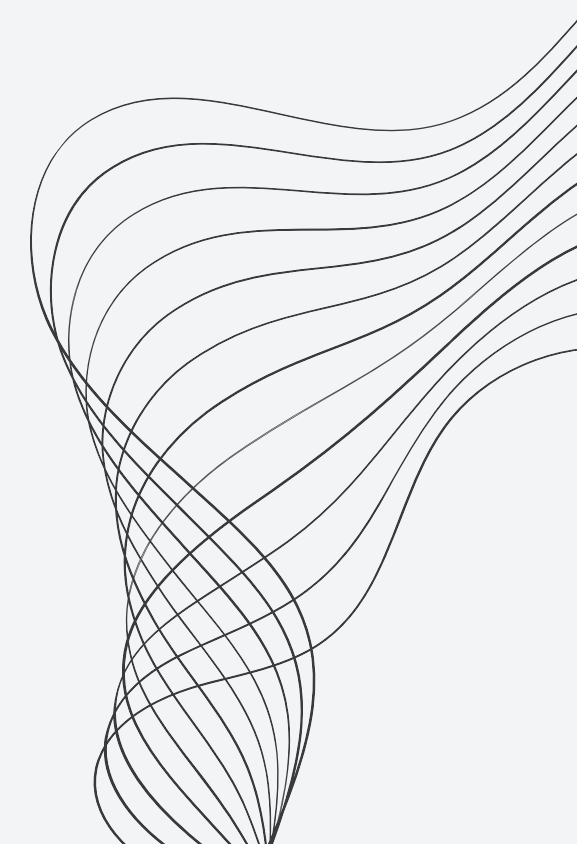


Importance of Reproducibility in Automated Workflows

- **Reliability:** Ensures that analyses and reports can be trusted by verifying that the same inputs yield the same results.
 - **Transparency:** Provides an audit trail for how data was processed, enabling traceability and simplifying troubleshooting.
 - **Collaborative Scalability:** Facilitates shared workflows across teams or organizations, as reproducible code and workflows allow anyone with the same data and environment to replicate the findings
- 

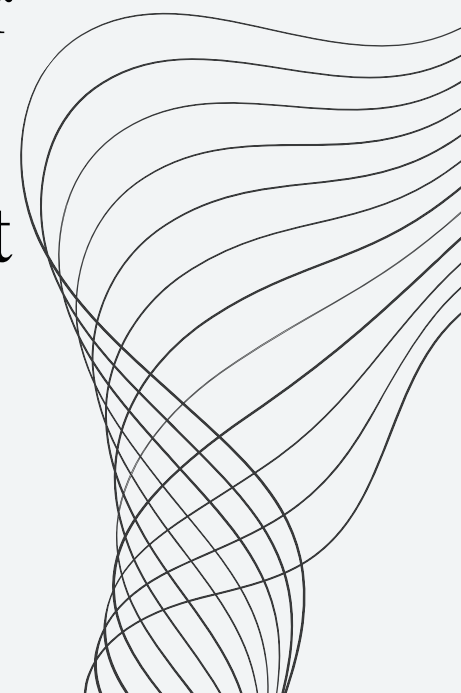


Examples of Automation in Data and Workflow Management

- ETL (Extract, Transform, Load): Automated data pipelines extract data from sources, transform it into a usable format, and load it into databases.
 - Scheduling and Monitoring: Tools like **Apache Airflow** and **Prefect** schedule and monitor data processing tasks.
 - Robotic Process Automation (RPA): Automates data entry, form filling, and other repetitive tasks across systems.
- 



Apache Airflow is an open-source workflow management tool primarily used to schedule, organize, and monitor data workflows

- **Pure Python:** Build workflows using Python, with flexible scheduling and task generation.
 - **User-Friendly UI:** Monitor and manage tasks through a modern, intuitive web interface.
 - **Robust Integrations:** Supports cloud services (GCP, AWS, Azure) and many other tools for easy integration.
 - **Easy to Use:** Python knowledge is enough to deploy complex workflows for ML, data transfer, etc.
 - **Open Source:** Contribute improvements easily, with active community support on Slack.
- 

DAGs

All 26 Active 10 Paused 16

Filter DAGs by tag

Search DAGs

<i>i</i> DAG	Owner	Runs <i>i</i>	Schedule	Last Run <i>i</i>	Recent Tasks <i>i</i>	Actions	Links
<input checked="" type="checkbox"/> example_bash_operator example example2	airflow	<div> <div>2</div> <div></div> <div></div> </div>	00***	2020-10-26, 21:08:11 <i>i</i>	<div> <div>6</div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>	▶ ↺ 🗑	...
<input checked="" type="checkbox"/> example_branch_dop_operator_v3 example	airflow	<div> <div></div> <div></div> <div></div> </div>	*/1****		<div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>	▶ ↺ 🗑	...
<input type="checkbox"/> example_branch_operator example example2	airflow	<div> <div></div> <div>1</div> <div></div> </div>	@daily	2020-10-23, 14:09:17 <i>i</i>	<div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div>11</div> </div>	▶ ↺ 🗑	...
<input checked="" type="checkbox"/> example_complex example example2 example3	airflow	<div> <div>1</div> <div>1</div> <div></div> </div>	None	2020-10-26, 21:08:04 <i>i</i>	<div> <div>37</div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>	▶ ↺ 🗑	...
<input checked="" type="checkbox"/> example_external_task_marker_child	airflow	<div> <div></div> <div>1</div> <div></div> </div>	None	2020-10-26, 21:07:33 <i>i</i>	<div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div>2</div> </div>	▶ ↺ 🗑	...
<input checked="" type="checkbox"/> example_external_task_marker_parent	airflow	<div> <div></div> <div>1</div> <div></div> </div>	None	2020-10-26, 21:08:34 <i>i</i>	<div> <div>1</div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>	▶ ↺ 🗑	...
<input checked="" type="checkbox"/> example_kubernetes_executor example example2	airflow	<div> <div></div> <div></div> <div></div> </div>	None		<div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>	▶ ↺ 🗑	...
<input checked="" type="checkbox"/> example_kubernetes_executor_config example3	airflow	<div> <div></div> <div>1</div> <div></div> </div>	None	2020-10-26, 21:07:40 <i>i</i>	<div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div>5</div> </div>	▶ ↺ 🗑	...
<input checked="" type="checkbox"/> example_nested_branch_dag example	airflow	<div> <div></div> <div>1</div> <div></div> </div>	@daily	2020-10-26, 21:07:37 <i>i</i>	<div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div>9</div> </div>	▶ ↺ 🗑	...
<input type="checkbox"/> example_passing_params_via_test_command example	airflow	<div> <div></div> <div></div> <div></div> </div>	*/1****		<div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> <div></div> </div>	▶ ↺ 🗑	...