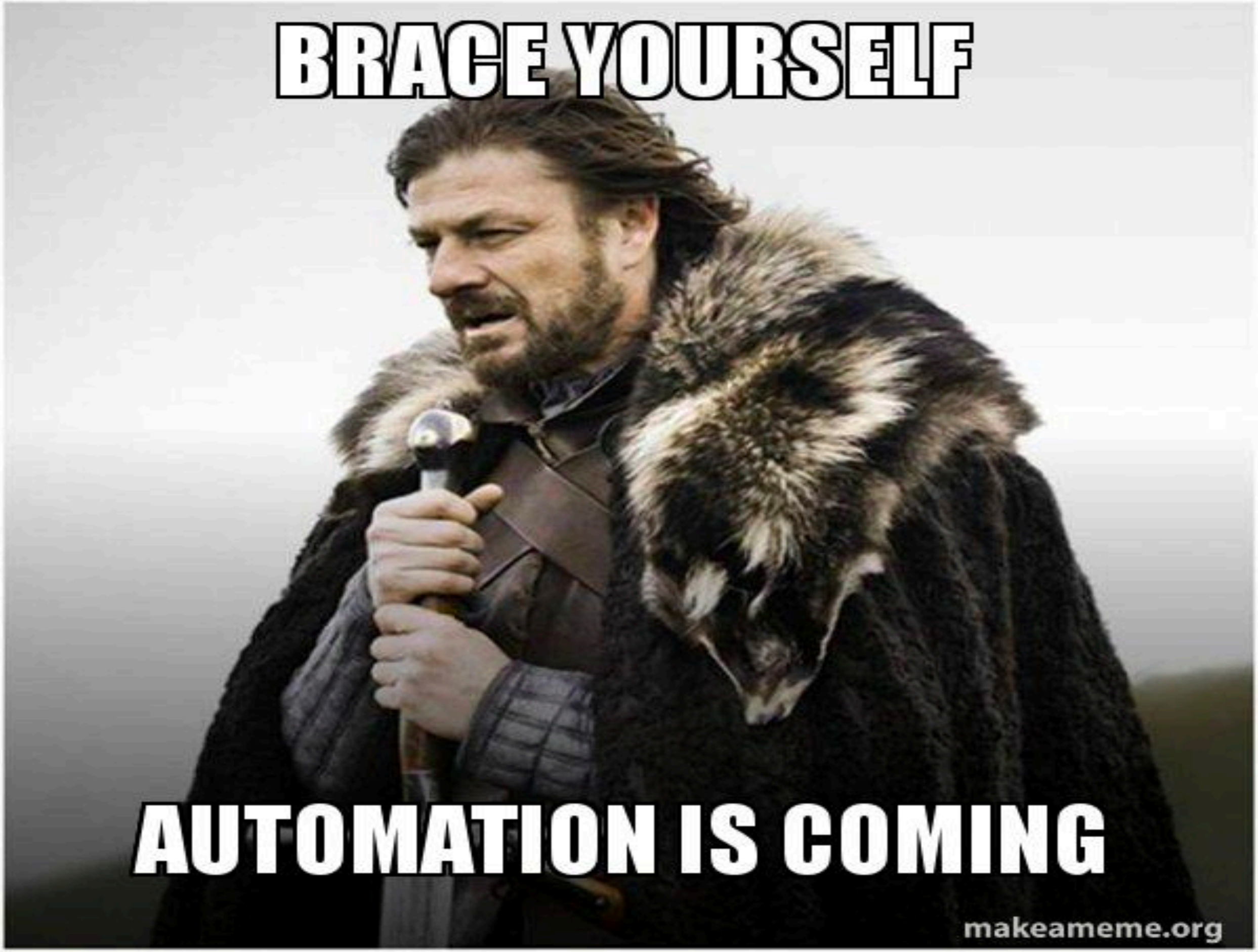

BRACE YOURSELF
AUTOMATION IS COMING
makeameme.org

Data Pipelines with Apache Airflow

Luther Hill
Data Scientist

Agenda

- Bio
- Data Science Truths
- Problem Statement
- Hands On
- Conclusion

Bio

BIO
I use data science to solve healthcare problems for the Army's Medical Department. I have been a Data Scientist for the last few years. Before that I worked as a data analytic manager for the Army.

Education
MS Computer Science: Data Analytic
BS Accounting

Interest

- Natural Language Processing
- Healthcare equity
- Diversity and inclusion

Data Science Truths

- Humans are more important than data.
- Your data is only as good as your story telling.
- Your data is biased and so are you.

Problem Statement

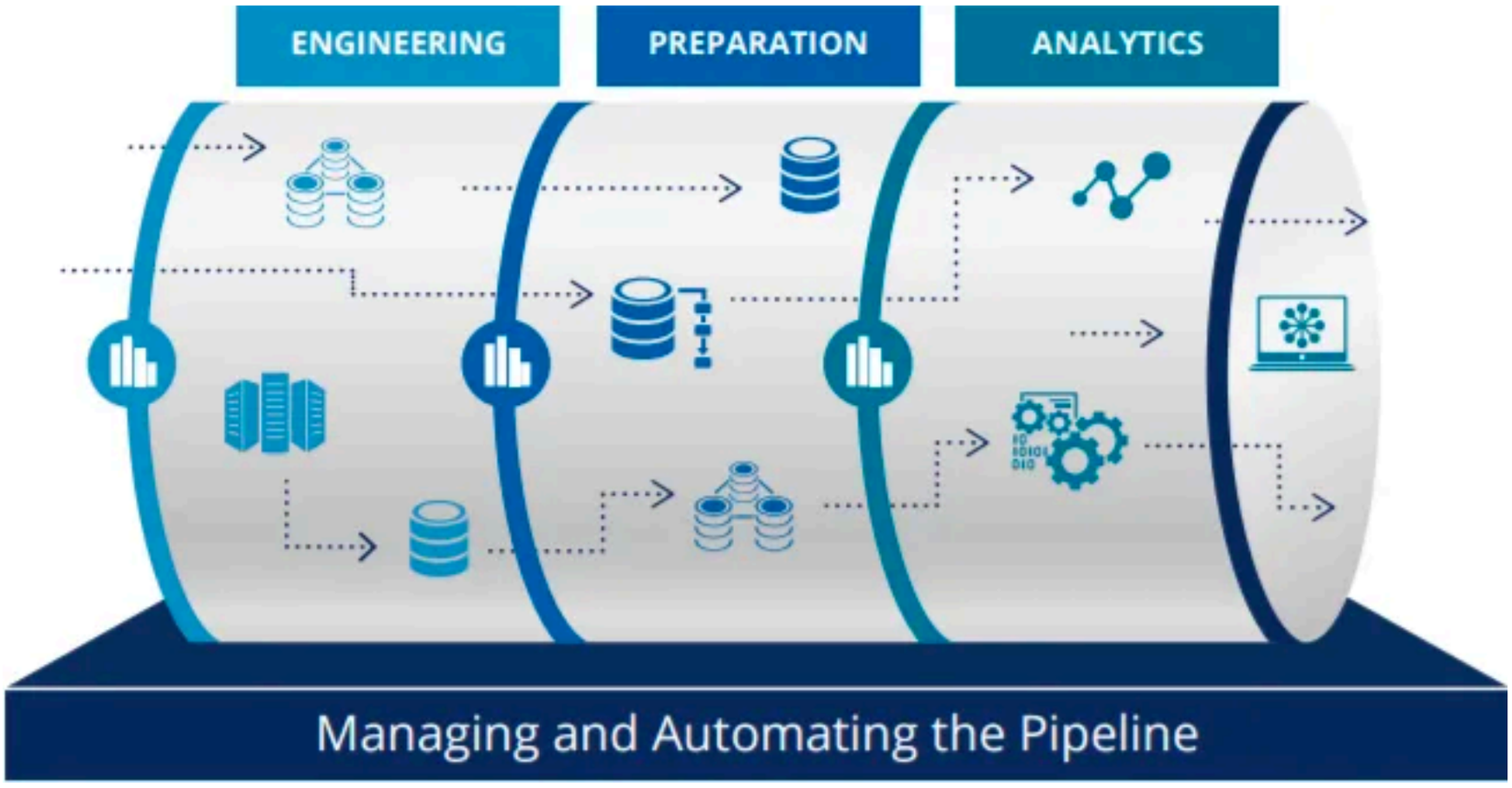
- We are a marketing firm tasked with finding out how people feel about our pycon2019.

Your Data is Dirty

unless proven otherwise

"It's in the database, so it's already good"

When working with data pipelines always remember these two statements.



Data Engineering

Why do pipelines matter?

- Analytics and batch processing is mission-critical as they power all data-intensive applications
- The complexity of the data sources and demands increase every day
- A lot of time is invested in writing, monitoring jobs, and troubleshooting issues.

This makes data engineering one of the most critical foundations of the whole analytics cycle.

Good data pipelines are:

- Reproducible: same code, same data, same environment -> same outcome
- Easy to productive: need minimal modifications from R&D to production

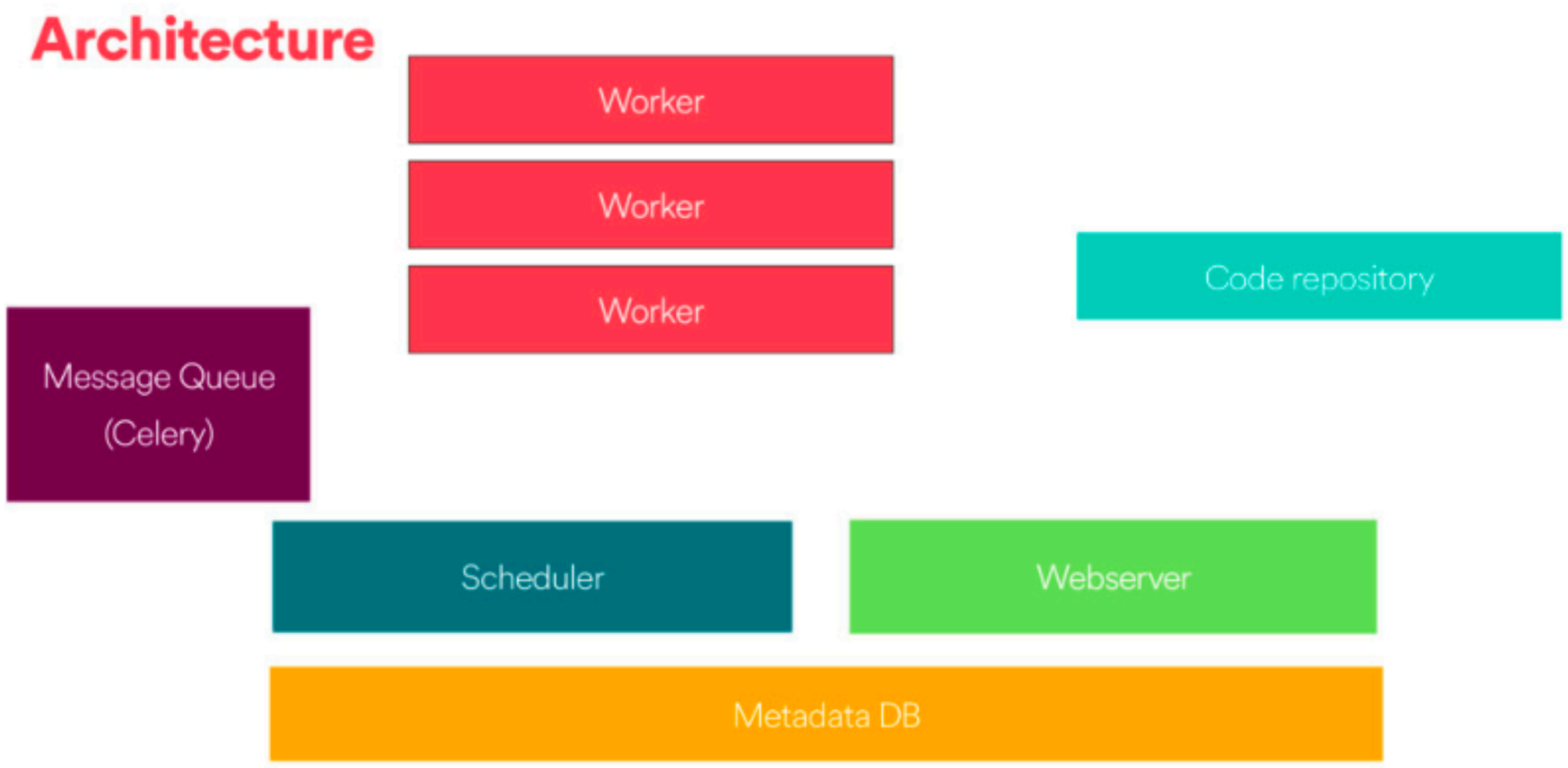
Apache Airflow

Basic Airflow concepts

- **Task**: a defined unit of work (operators)
- **Task instance**: an individual run of a single task. indicative state(“running”, “success”, “failed”, “skipped”, “up for retry”)
- **DAG**: Directed acyclic graph, a set of tasks with explicit execution order, beginning, and end
- **DAG run**: individual execution/run of a DAG

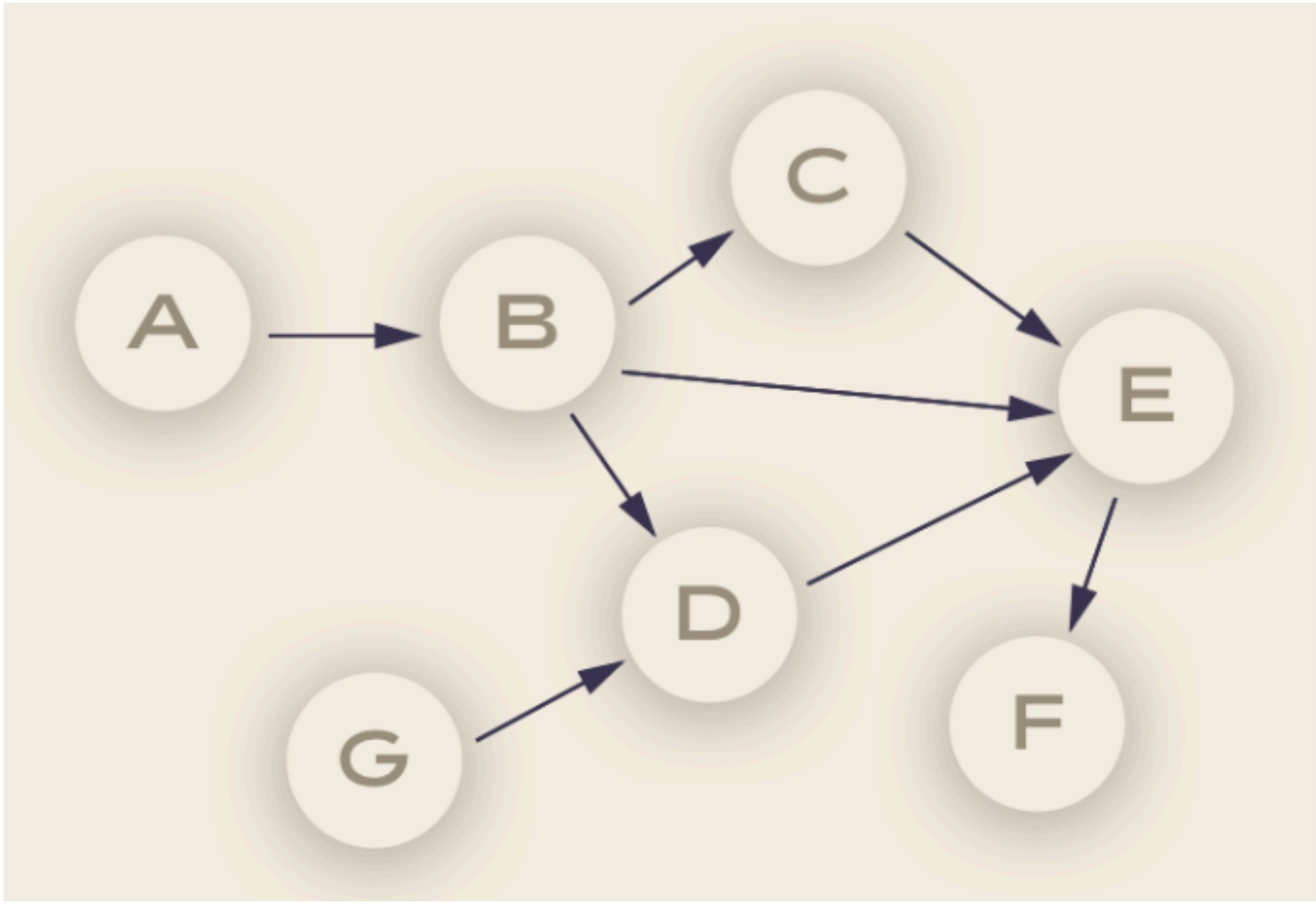
Apache Airflow

- Workflow as code
- Scheduler
- Executor
- Metadata database
- Defining Task



Apache Airflow

Architecture



Debunking the DAG

The vertices and edges (the arrows linking the nodes) have an order and direction associated to them

Hands On

- Apache Airflow
- MSQL
- NIX system
- Twitter developer account