

# Self-Intro

#### Data Engineering Consultant

#### Tools

• Python, AWS, Airflow, Ansible

**Business Problems** 

- Batch Processing Workflows ELT / ETL
- Ground-Up Data Infrastructures







## Maslow at the Blackfoot Reservation in 1938















### But why Another mental model for data



Focus is on fundamentals

Reasoning > Principles > Tools



01.

#### Automation





#### 💁 Casey @kccarrell · Feb 5

Someone sent me an excel file to fix today because it kept 'blowing up'. It had 20 hidden sheets, 100s of bespoke protections, dozens of named ranges, half of which had defunct references, and a bunch of trash VBA. I told him without sarcasm that it would be easier to start over.



### Without Automation





What is a good baseline for Automation?





What is a good baseline for Automation?

#### Scripts

- Source control and Schedule below scripts
  - Script Existing Manual and Predictable Data Wrangling
  - Move legacy click and drag workflows over to scripts





### What does robust Automation look like?

More layers of complexity





### What does robust Automation look like?

More layers of complexity

• Infrastructure as Code (IaC)





### What does robust Automation look like?

More layers of complexity

- Infrastructure as Code (IaC)
- Data Workflow Orchestration





### Why Airflow? It's Extensible

**Engineering Talent** 

• Leverages Python language as the analytics standard

Technical

- Connections to any data source
- Lightweight backend works on any Linux/Unix Server
- Code as Abstraction Layer









#### Extract





### Extract (v.)





#### Extract (v.)

Without Extraction, there are no ingredients for which our analysts to do their work

Without ingredients, any optimization is premature.





#### Extract (v.)

Either no-code Data Integration SaaS solution

#### Or

Fully automate your Data Source connections in code









#### Cheaper storage killed ETL.

And ELT took its place.





#### Load

#### Data Lakes

- Raw data will be in a rough state.
- Cloud Storage allows Analysts to query
  - Queries may be complex





#### Load

#### Data Lakes

- Raw data will be in a rough state.
- Cloud Storage allows Analysts to query
  - Queries may be complex
- Daily Snapshots (more info)





#### Load

#### Data Lakes

- Raw data will be in a rough state.
- Cloud Storage allows Analysts to query
  - Queries may be complex
- Daily Snapshots (more info)
- Optimize with Parquet files







04.

### Transform





## Transform





#### Data Work that can be kept in SQL only.

Why?

### Transform





1. <u>Maintainable Workflows</u>





- 1. <u>Maintainable Workflows</u>
- 2. More Complexity
  - a. Remove Data Silos





- 1. <u>Maintainable Workflows</u>
- 2. More Complexity
  - a. Remove Data Silos
  - b. Parameterize your SQL





### Parameterize your SQL

#### SELECT {{ cols }} FROM tbl {{ where }}





- 1. <u>Maintainable Workflows</u>
- 2. More Complexity
  - a. Remove Data Silos
  - b. Parameterize your SQL
  - c. Data Quality Testing



05.

## Optimize Analysis









Time Sensitive Reporting

• Spark





Time Sensitive Reporting

Spark

Custom Data Transformations

• Jupyter Notebooks





Time Sensitive Reporting

• Spark

Custom Data Transformations

- Jupyter Notebooks Large Scale Processes
  - Reduce Computational Cost with Systems Engineering



06.

## Machine Learning



## Machine Learning





## Machine Learning













#### Streaming for Data Analysis, alone, is rare.

### Streaming



#### Conclusion Big "Why?"s







#### Transparency And Reproducibility



#### Enabling Ethics



## Thank you!

Say hi! Ask questions!



Writing: angelddaz.substack.com Contact: angel@ocelotdata.com

