Data Engineering
Hierarchy of Needs
Angel D’az
Self-Intro

Data Engineering Consultant

Tools
- Python, AWS, Airflow, Ansible

Business Problems
- Batch Processing Workflows ELT / ETL
- Ground-Up Data Infrastructures
Physiological needs: food, water, warmth, rest

Safety needs: security, safety

Belongingness and love needs: intimate relationships, friends

Esteem needs: prestige and feeling of accomplishment

Self-actualization: achieving one’s full potential, including creative activities

Self-fulfillment needs
Maslow at the Blackfoot Reservation in 1938
Self-actualization:
achieving one’s full potential, including creative activities

Esteem needs:
prestige and feeling of accomplishment

Belongingness and love needs:
immediate relationships, friends

Safety needs:
security, safety

Physiological needs:
food, water, warmth, rest

Self-fulfillment
Psychological
Basic
needs
needs
needs
THE DATA SCIENCE HIERARCHY OF NEEDS

LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT

AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
THE DATA SCIENCE HIERARCHY OF NEEDS

COLLECT

MOVE/STORE

EXPLORE/TRANSFORM

AGGREGATE/LABEL

LEARN/OPTIMIZE

AI, DEEP LEARNING

A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS

ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
Automate/Orchestrating
Extract
Load
Transform
Optimize Analysis
ML/Models
streaming
But why Another mental model for data
Focus is on fundamentals

Reasoning > Principles > Tools
01.

Automation
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.
Why Automation first?
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
02.

Extract
Extract (v.)
Without Extraction, there are no ingredients for which our analysts to do their work

Without ingredients, any optimization is premature.
Either no-code Data Integration SaaS solution

Or

Fully automate your Data Source connections in code
03. Load
Cheaper storage killed ETL.

And ELT took its place.
Data Lakes
- Raw data will be in a rough state.
- Cloud Storage allows Analysts to query
  - Queries may be complex
Data Lakes

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

1. Maintainable Workflows
Why SQL only?

1. Maintainable Workflows
2. More Complexity
   a. Remove Data Silos
Why SQL only?

1. Maintainable Workflows
2. More Complexity
   a. Remove Data Silos
   b. Parameterize your SQL
Parameterize your SQL

```
SELECT {{ cols }}
FROM tbl
{{ where }}
```
Why SQL only?

1. **Maintainable Workflows**
2. **More Complexity**
   a. Remove Data Silos
   b. Parameterize your SQL
   c. Data Quality Testing
05. Optimize Analysis
Optimize Analysis
Optimize Analysis

Time Sensitive Reporting
- Spark
Optimize Analysis

Time Sensitive Reporting
- Spark
Custom Data Transformations
- Jupyter Notebooks
Optimize Analysis

- Time Sensitive Reporting
  - Spark
- Custom Data Transformations
  - Jupyter Notebooks
- Large Scale Processes
  - Reduce Computational Cost with Systems Engineering
06.

Machine Learning
Machine Learning
07.

Streaming
Streaming for Data Analysis, alone, is rare.
Conclusion
Big “Why?”s
Transparency And Reproducibility
Enabling Ethics
Thank you!

Say hi! Ask questions!

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