Data Engineering

Hierarchy of Needs

Angel D'az

Self-Intro

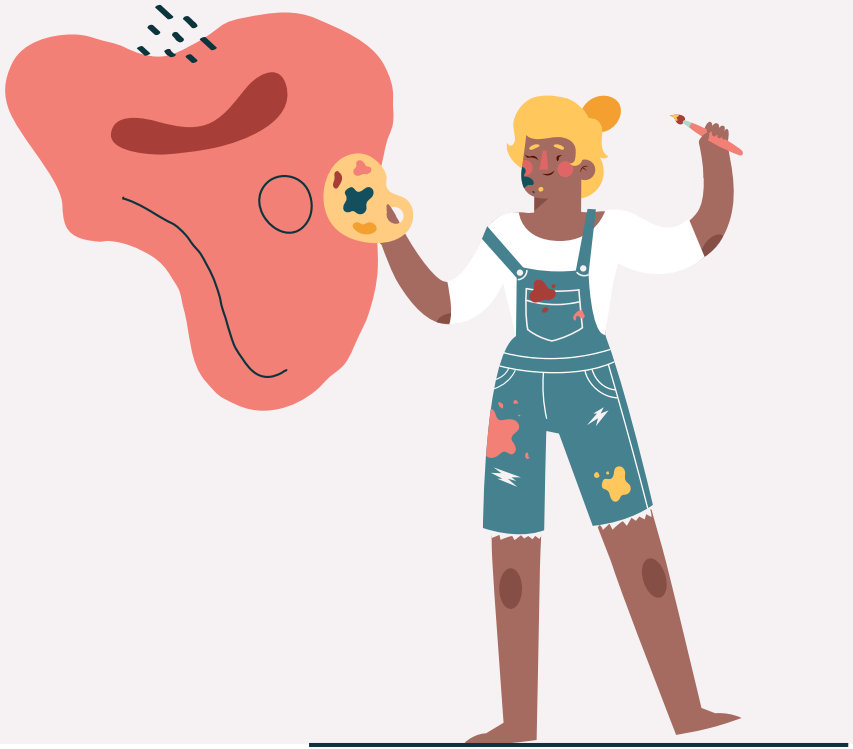
Data Engineering Consultant

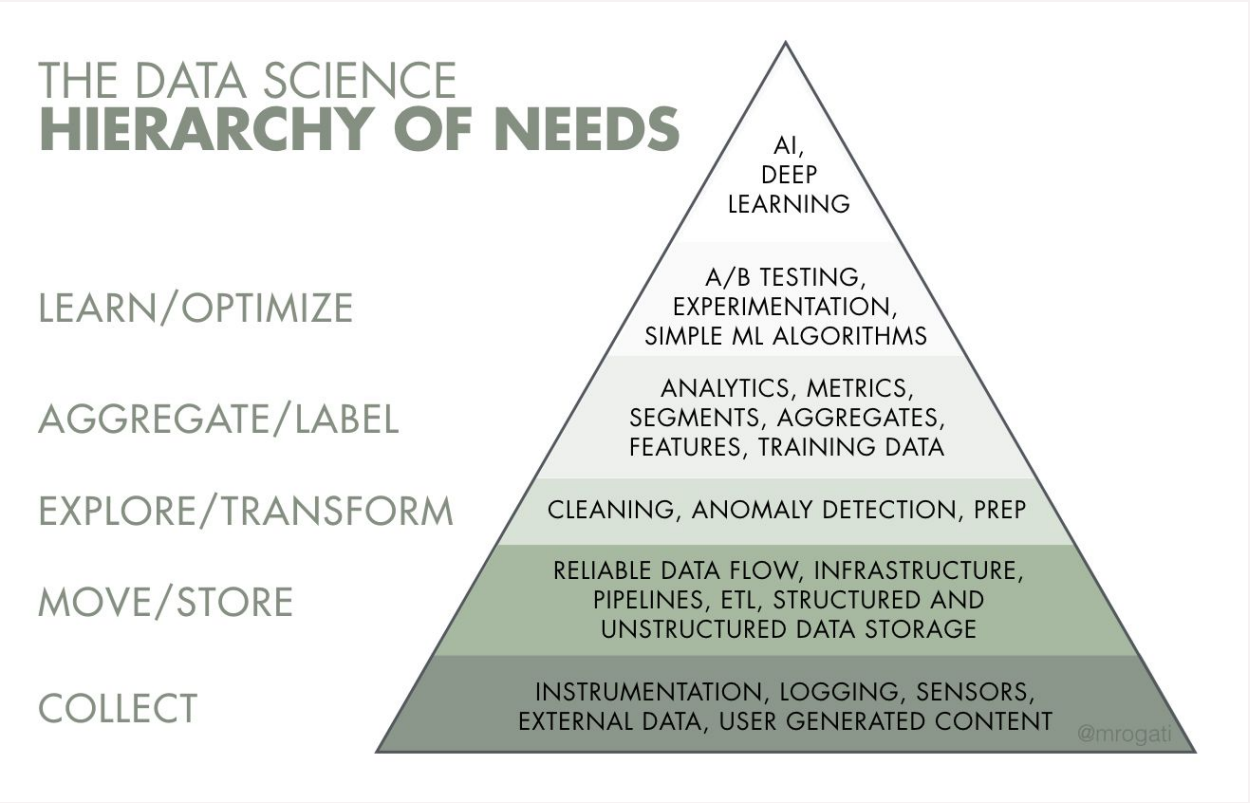
Tools

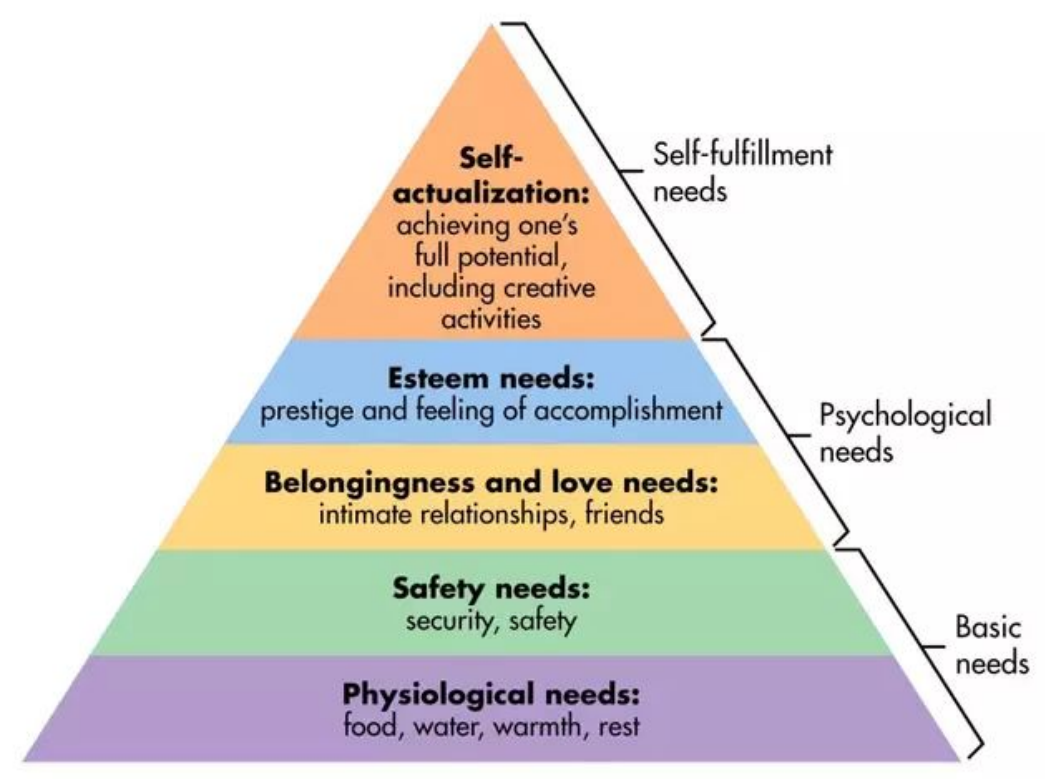
- Python, AWS, Airflow, Ansible

Business Problems

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

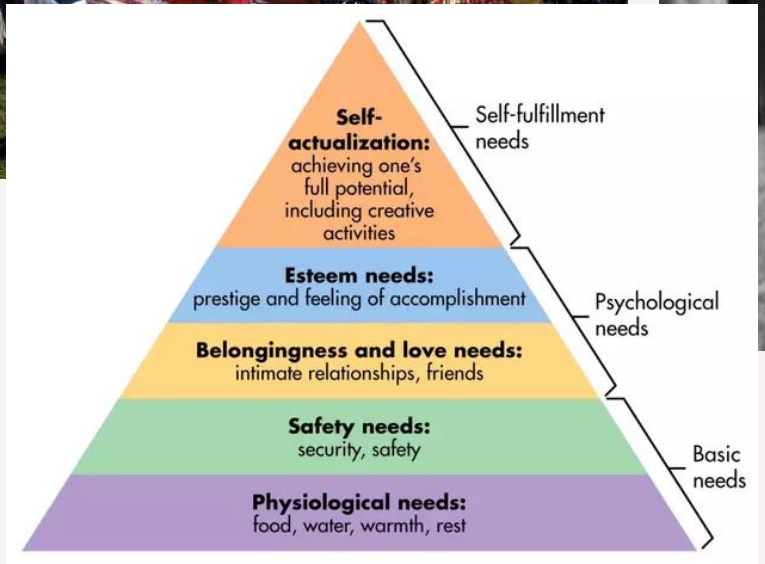



THE DATA SCIENCE
HIERARCHY OF NEEDS
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI,
DEEP
LEARNING
A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS
ANALYTICS, METRICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE,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

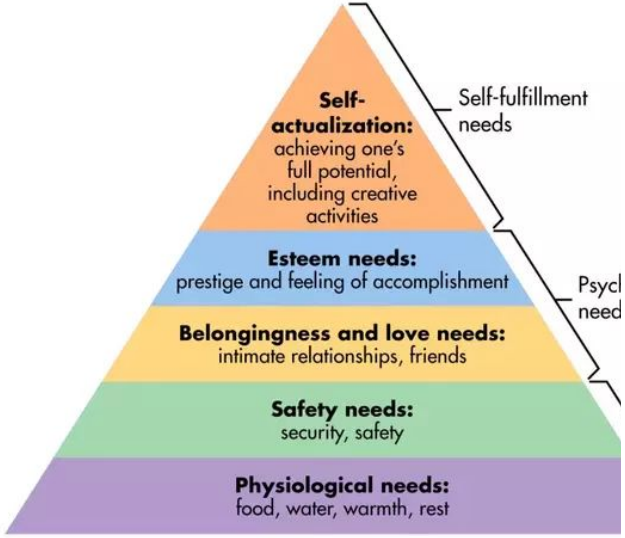

Self-actualization: achieving one's full potential, including creative activities
Esteem needs: prestige and feeling of accomplishment
Belongingness and love needs: intimate relationships, friends
Safety needs: security, safety
Physiological needs: food, water, warmth, rest
Self-fulfillment needs
Psychological needs
Basic needs

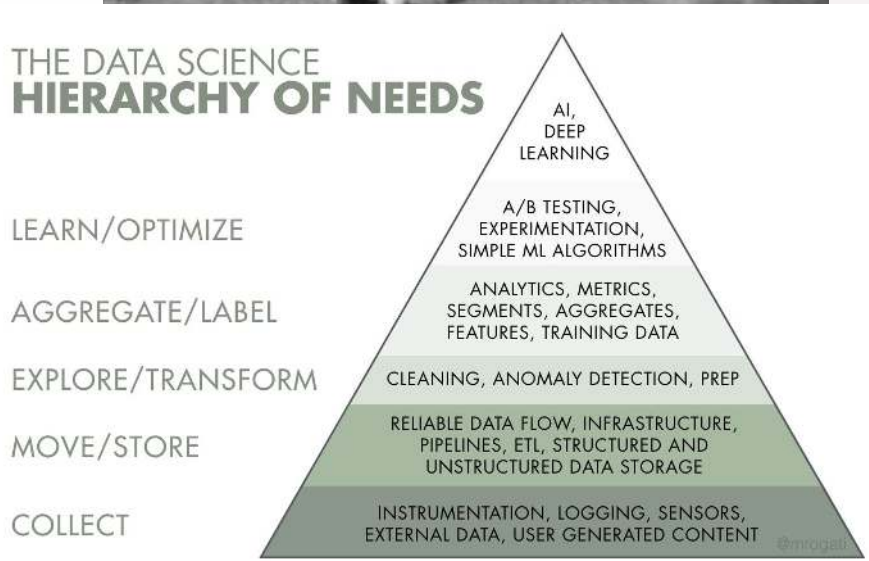
Maslow at the Blackfoot Reservation in 1938

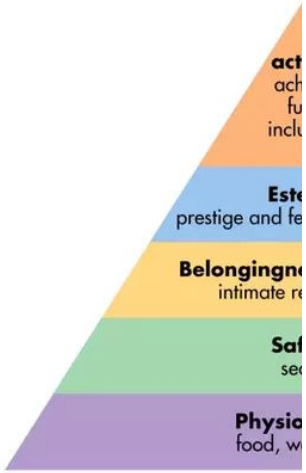

Glenbow Archives PA-3385-185


Self-actualization: achieving one's full potential, including creative activities
Esteem needs: prestige and feeling of accomplishment
Belongingness and love needs: intimate relationships, friends
Safety needs: security, safety
Physiological needs: food, water, warmth, rest
Self-fulfillment needs
Psychological needs
Basic needs

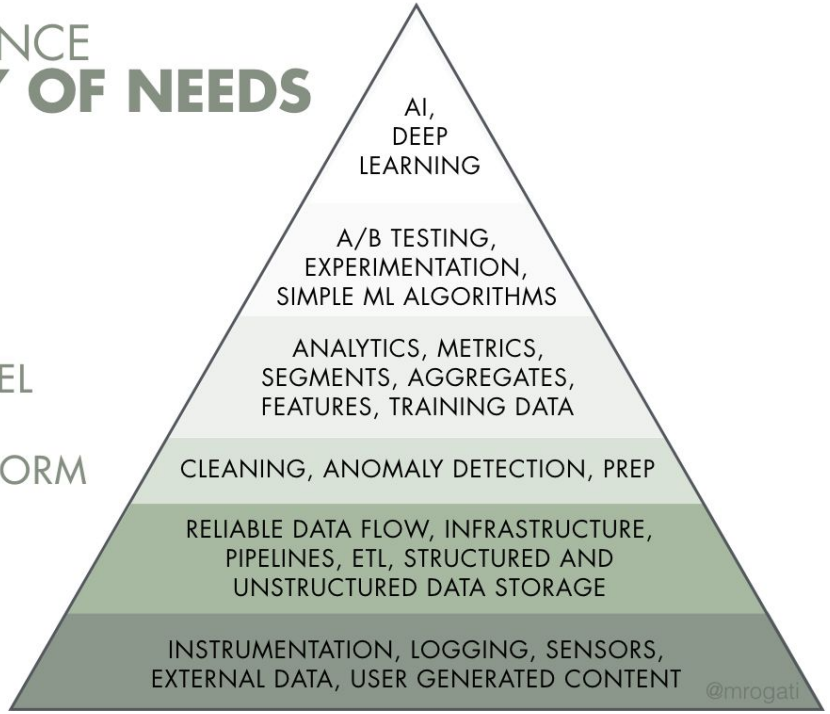

Glenbow Archives PA-3385-185


Self-actualization: achieving one's full potential, including creative activities
Esteem needs: prestige and feeling of accomplishment
Belongingness and love needs: intimate relationships, friends
Safety needs: security, safety
Physiological needs: food, water, warmth, rest
Self-fulfillment needs
Psycl
need


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



THE DATA SCIENCE HIERARCHY OF NEEDS



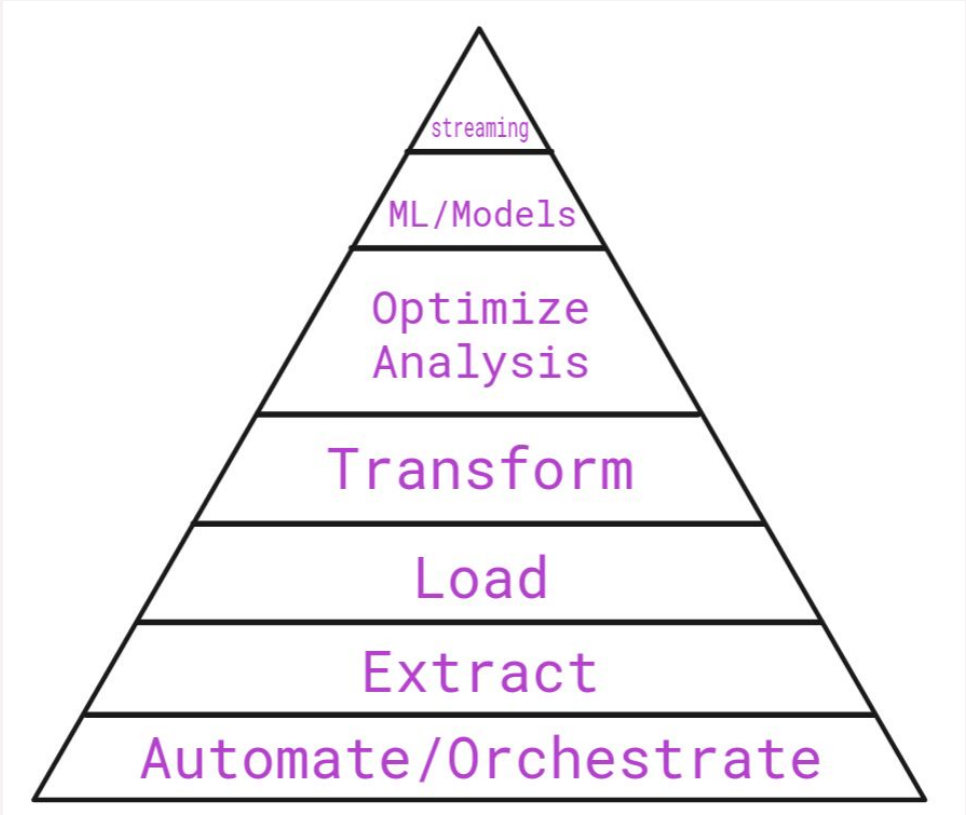
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT


streaming
ML/Models
Optimize
Analysis
Transform
Load
Extract
Automate/Orchestrate

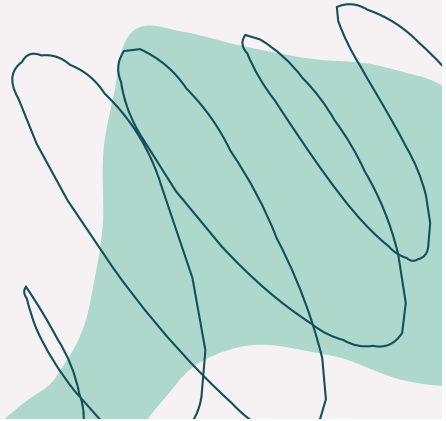
But why Another mental model for data

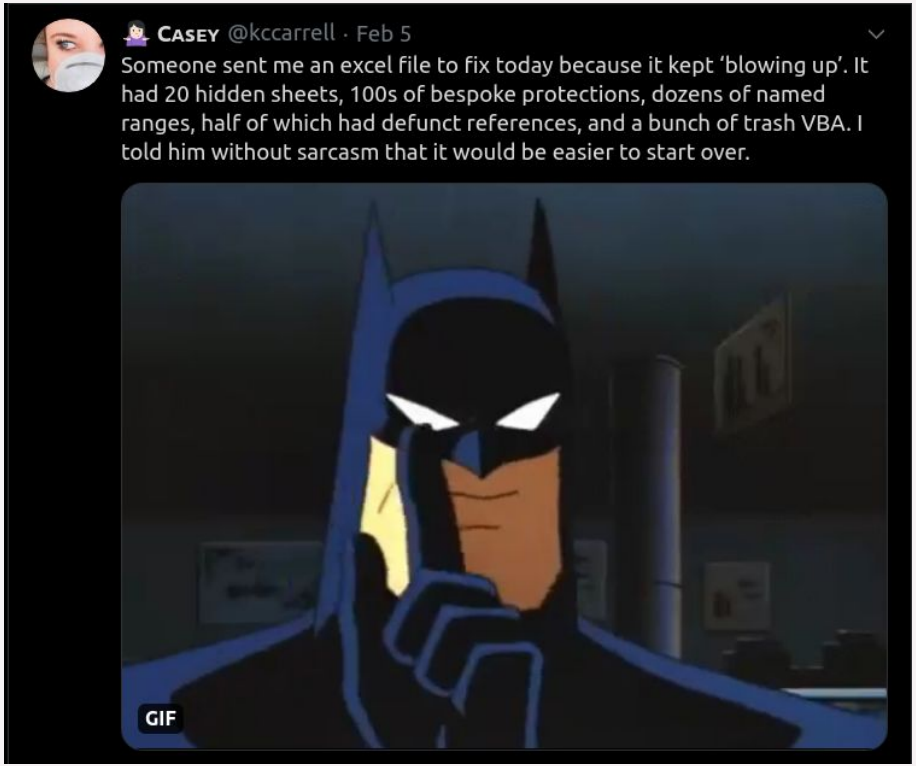
Focus is on fundamentals

Reasoning > Principles > Tools

01.

Automation





Without Automation

Why Automation first?

What is a good baseline for Automation?

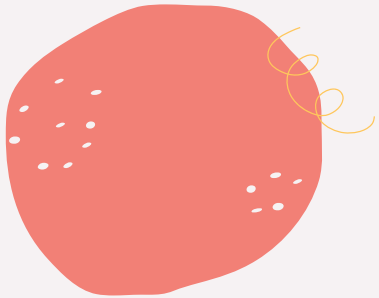
What is a good baseline for Automation?

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?

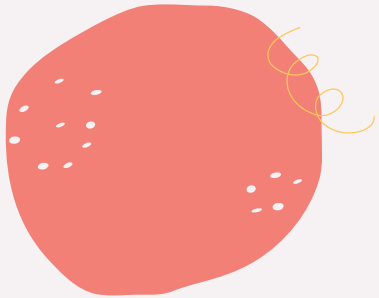
More layers of complexity

- 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.

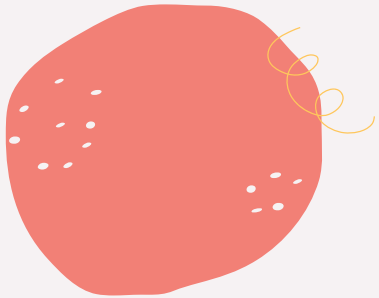
Extract

Extract (v.)

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

03.

Load

Load

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)

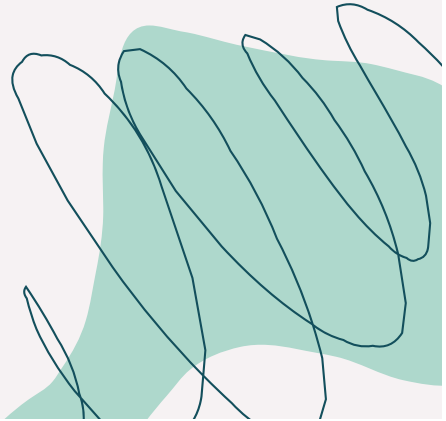
Load

Data Lakes

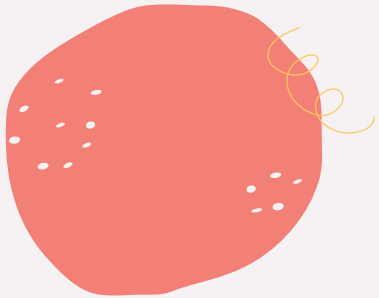
- Raw data will be in a rough state.
- Cloud Storage allows Analysts to query
  - Queries may be complex
- Daily Snapshots (more info)
- Optimize with Parquet files

04.

Transform



Transform



Transform

Data Work that can be kept in SQL only.

Why?

Why SQL only?

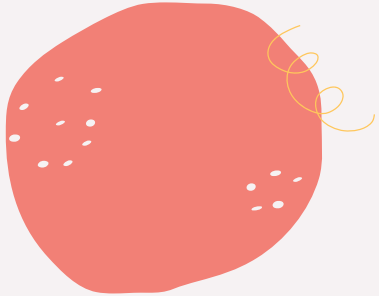
1. Maintainable Workflows

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

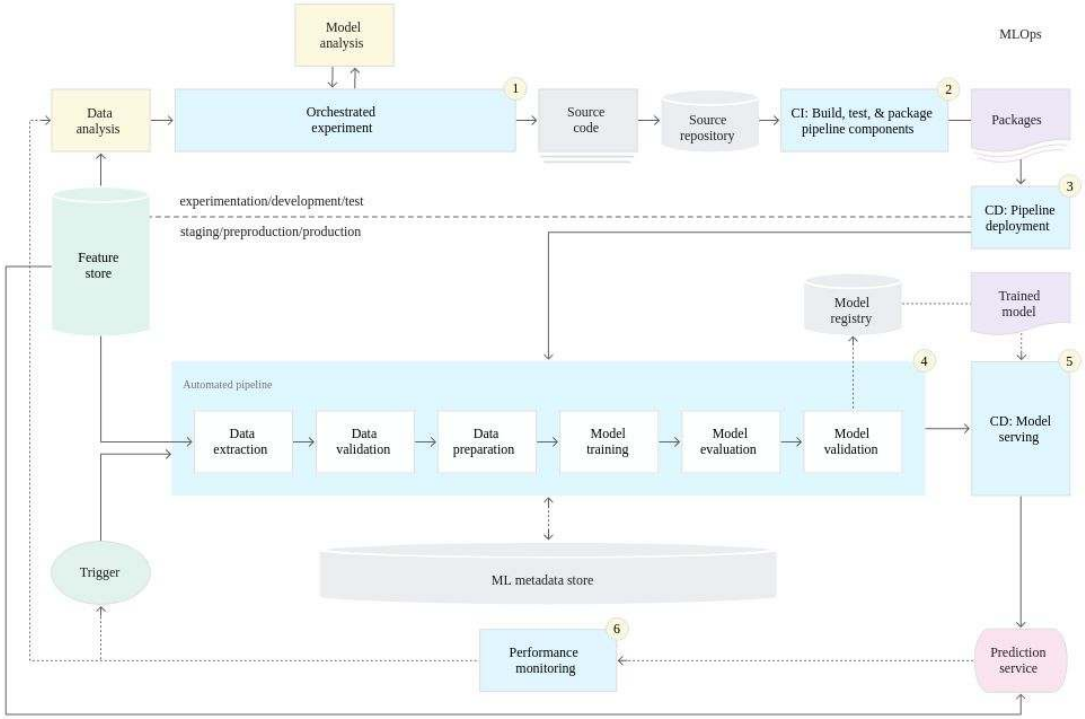
Custom Data Transformations
- Jupyter Notebooks

Large Scale Processes
- Reduce Computational Cost with Systems Engineering

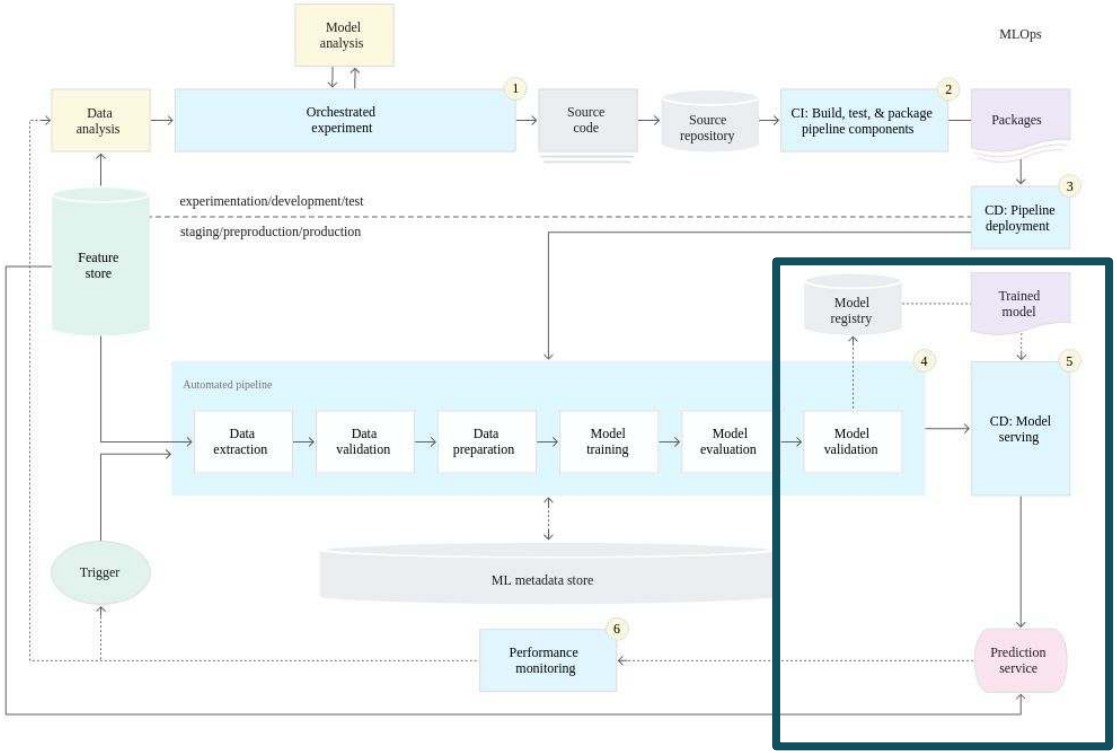
06.

Machine Learning

Machine Learning



Machine Learning

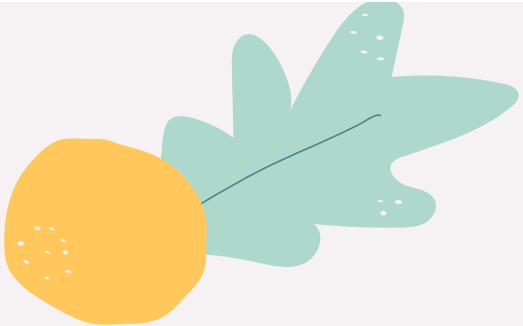


07.

Streaming

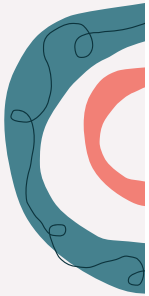
Streaming

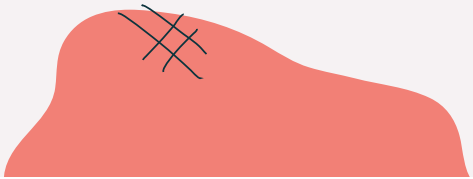
Streaming for Data Analysis, alone, is rare.



Conclusion

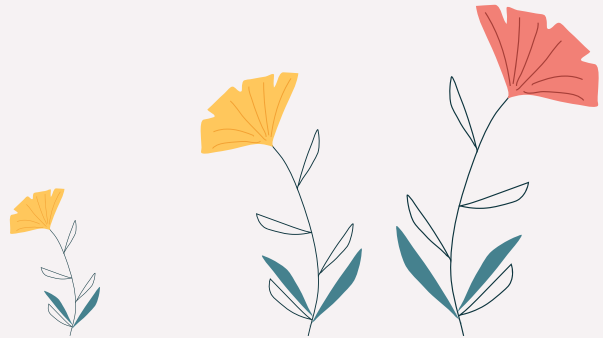
Big “Why?”s

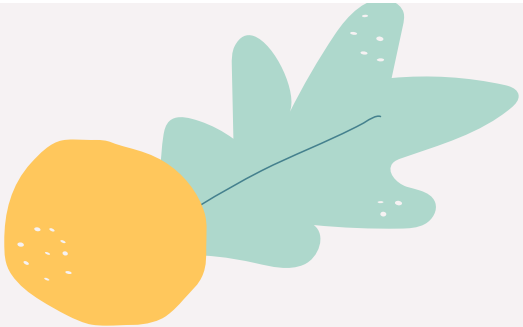




Transparency And Reproducibility

Enabling Ethics





Thank you!

Say hi! Ask questions!

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