Los Alamos NATIONAL LABORATORY

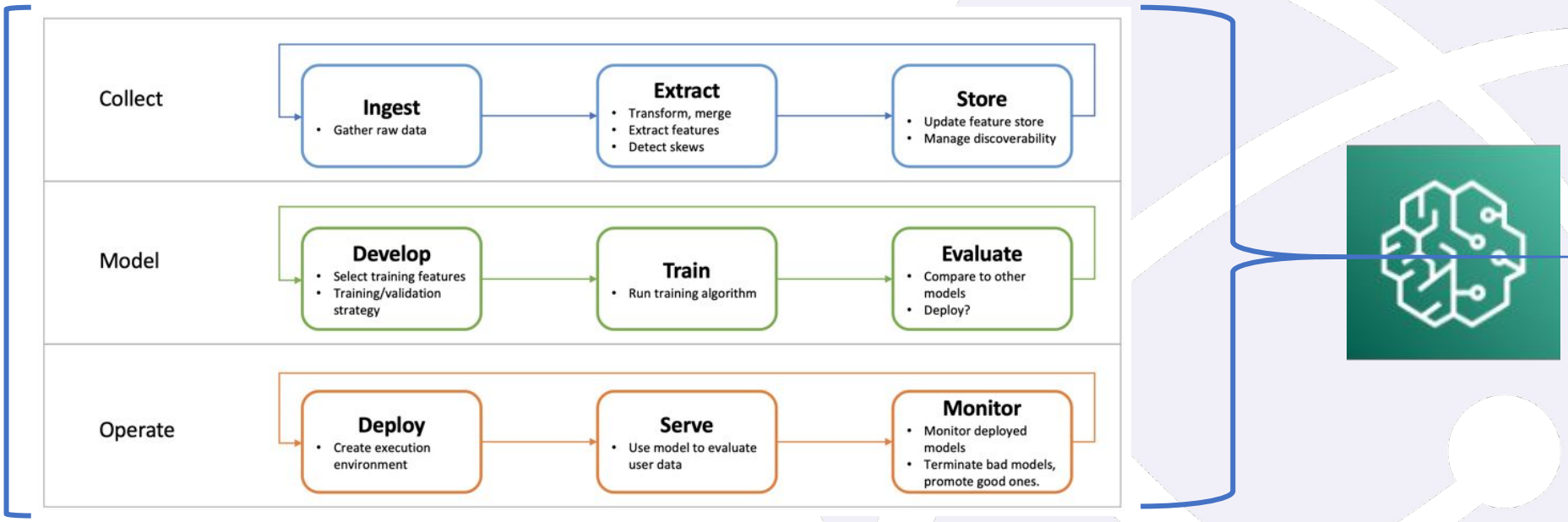
AWS SageMaker for Machine Learning Operations at LANL

William Rosenberger (A-1)
Ashlynn Daughton (A-1)
James Wernicke (A-4)

09/14/2021

LA-UR-21-28920

Data Operations

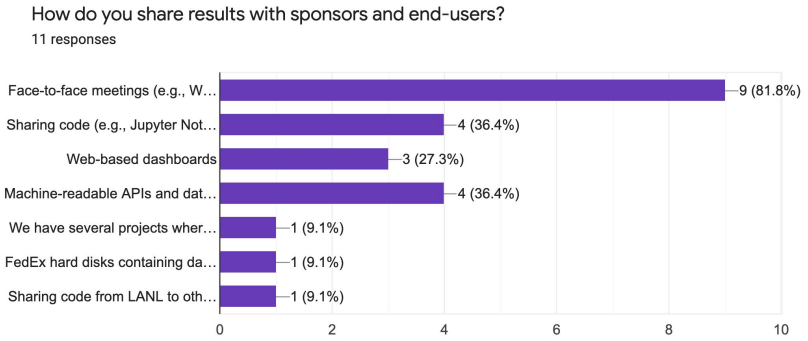


Approach

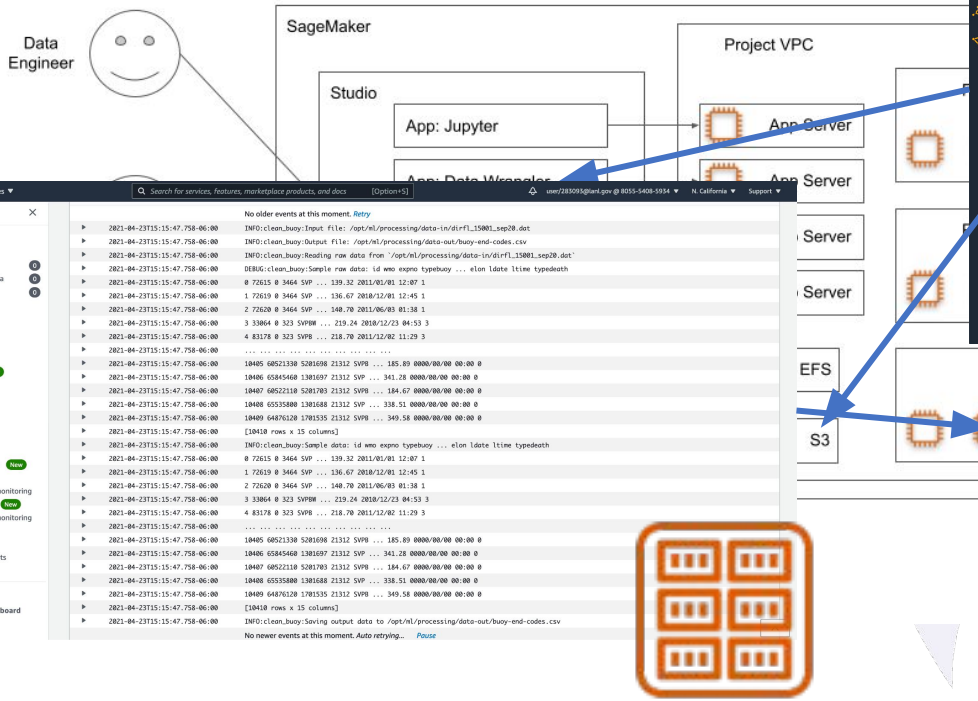
- Gather requirements
- Learn SageMaker
- Evaluate SageMaker

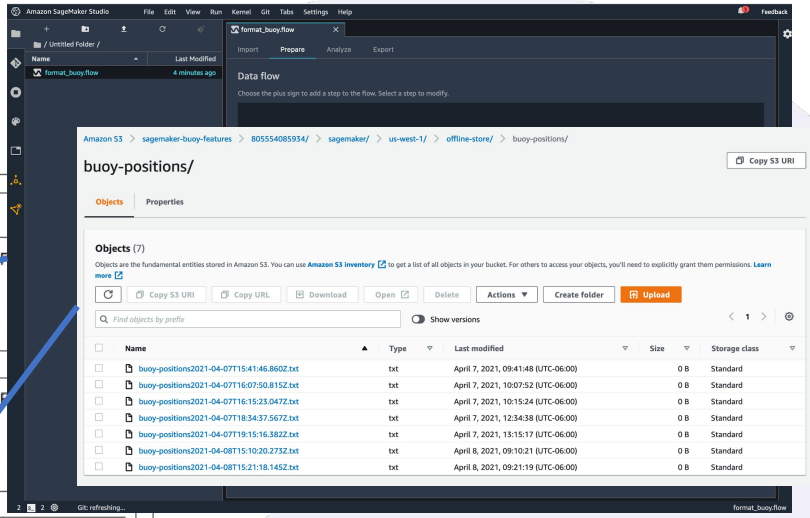
Requirements Elicitation

- **Gather requirements**
  - Created a poll
  - Distributed to SMEs throughout the lab
  - 11 respondents



SageMaker Architecture





Results

- Abstracts away the complexity of AWS
  - SageMaker does a good job of integrating with a large number of AWS services in a transparent manner
  - Most development tasks are controlled through a familiar Jupyter Notebook interface
- Cost
  - Complex services imply complex pricing model
  - SageMaker prioritizes usability over minimizing costs (large instances by default, automatic startup of new instances)

Results

- Use cases
  - New development teams and environments that have little existing computational infrastructure
  - Existing teams that need to transition from medium-sized data sets to larger-than-memory data sets
  - Projects that need to coordinate the execution of multiple independent code bases

Results

- Learning curve
  - LANL has significant momentum built up around existing HPC-based infrastructure
    - SageMaker requires learning new best practices and usage patterns
  - SageMaker UI does a good job of helping users learn about advanced features
  - Projects may need assistance from AWS support in order to use SageMaker effectively

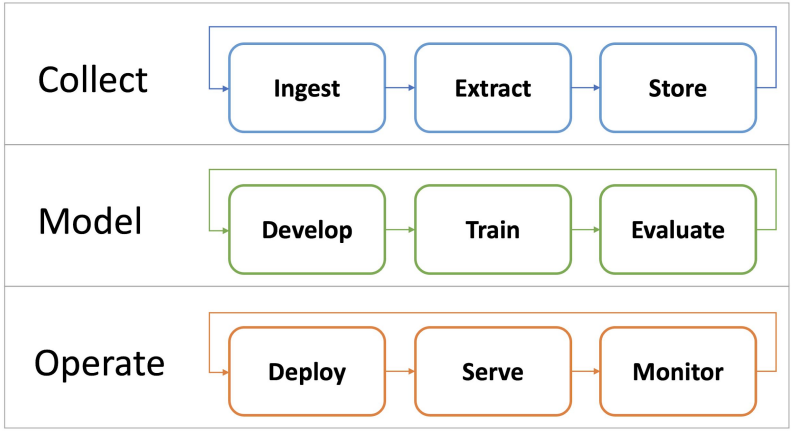
Results

- Lessons learned from SageMaker
  - Power of having a centralized feature store
  - Containers provide powerful flexibility
  - Decouple training code from serving code
  - Jupyter Notebooks make a *fantastic* learning environment

Thank you!

- ISTI IPD for funding
- LANL Cloud Services Team (NIE) for excellent support
- Alice Barthel (T-3) for pointing us to example data sets

AWS SageMaker for Machine Learning Operations at LANL



Operating complex data management pipelines for data science and machine learning requires complex systems managed by expert data engineers and scientists. Projects seeking to provide long-term results must invest heavily in systems to support the complexity of bringing a model to production. Services like AWS SageMaker provide an off-the-shelf solution to this problem, helping to avoid duplicate work between data science projects.

Project Description

Building end-to-end data science and machine learning pipelines requires significant investment in complex data management solutions. Can SageMaker simplify this problem?

Project Outcomes

- ***SageMaker is a powerful system with extensive tooling***
- ***Projects should be aware of cost and learning curve***
- ***Significant lessons can drawn from the SageMaker architecture***

PI: William Rosenberger
Total Project Budget: $20,750
ISTI Focus Area: Data Science and Artificial Intelligence

END