Course Info: OMSBA 5315 – Big Data Analytics

Date: March 14, 2025

Name: Shreeya Sampat

Title: Airflow in Action: Supercharging Data Science and ML Workflows at Apple

Source: https://www.astronomer.io/blog/airflow-in-action-apple/

Introduction:

Apple is using Apache Airflow to improve how it manages machine learning (ML) workflows. Data scientists often struggle with manual processes, debugging issues, and handling large-scale workflows. These challenges slow down progress and make it harder to deploy ML models efficiently. As mentioned in the *Astronomer blog*, these inefficiencies create bottlenecks in production.

To fix this, Apple enhanced Airflow's Papermill operator, allowing Jupyter Notebooks to run in multiple programming languages like Python and Scala. According to *Apple Machine Learning Research*, these improvements make workflows faster, more manageable, and scalable using cloud-based tools like Kubernetes. The goal is to bridge the gap between testing models and using them in real-world applications. Now, Apple's data teams can experiment, test, and deploy ML models with less manual effort, improving overall efficiency.

Analysis of Three Elements:

- Actor: Apple's data scientists, ML engineers, and platform teams develop and maintain machine learning models. Apple users also play a role by generating data through searches, purchases, and reviews, which help improve AI-driven recommendations.
- System: The system includes Apache Airflow for workflow automation, Papermill for executing Jupyter Notebooks, Spark and Scala for processing large data, and Kubernetes for cloud-based execution. These tools work together to make ML workflows faster and more efficient.
- Goal: The goal is to automate and optimize ML workflows, reducing manual work and improving scalability. This ensures faster processing, better AI-driven recommendations, and an improved user experience.

Analysis of Additional Elements:

- Stakeholders: Apple's data teams benefit from automation, business teams gain better insights, and Apple users receive improved recommendations and personalized services.
- Preconditions: Airflow, Jupyter, Spark, and Kubernetes must be configured, cloud resources must be available, and relevant user data must be collected for models to function properly.
- Triggers: The system activates when new ML models need testing, scheduled jobs run, or user interactions generate new data, ensuring continuous improvements.

Implementations / Technical Foundations:

Apple's system relies on Apache Airflow for automation. The company made key improvements to the Papermill operator, making it compatible with multiple programming languages. According to the *Apache Airflow Documentation*, Apple also optimized it to run Jupyter Notebooks remotely on Kubernetes, making large-scale workflows more efficient.

Cloud scalability is another important aspect. Instead of running everything locally, Apple now executes workflows in cloud environments, improving flexibility and speed. Debugging is also much easier. As mentioned in *Astronomer*, detailed logs track every workflow step, allowing engineers to quickly fix problems.

Implications on the case in terms of Descriptive / Predictive / Prescriptive Analytics:

- Descriptive Analytics: Apple improves descriptive analytics by tracking and logging workflow performance. With detailed records, debugging becomes easier, and engineers can quickly find and fix issues. As stated in *Astronomer*, this makes the system more reliable and efficient.
- Predictive Analytics: Apple uses predictive analytics in several areas, including sentiment analysis of iPhone customer reviews. By studying user feedback, the system improves model accuracy and performance. According to *Apple Machine Learning Research*, predictive analytics helps Apple refine its AI models, making them more effective.
- Prescriptive Analytics: This system also supports prescriptive analytics by automating decision-making. By allowing Jupyter Notebooks to execute tasks automatically, it reduces manual intervention. As mentioned in *Astronomer*, this makes Apple's workflows faster and ensures that ML models continuously improve.

Your Evaluations / Recommendations for Future Applications:

Apple's improvements to Airflow and Jupyter Notebooks have made ML workflows much more efficient. The ability to support multiple programming languages, cloud execution, and better monitoring has streamlined processes for data scientists. As highlighted in *Shrestha's slides*, Apple has successfully scaled its AI and ML workflows to handle massive datasets.

For future improvements, Apple could expand event-driven workflows. This would allow models to update instantly based on real-time user activity. Another improvement would be deeper integration with Apple's broader data ecosystem, making AI-driven features even smarter. As suggested in *Apple Machine Learning Research*, Apple could also incorporate additional ML tracking tools like MLflow or TensorBoard to improve model monitoring and management.

References - External Sources:

1. Astronomer. (2024). *Airflow in Action: Supercharging Data Science and ML Workflows at Apple*. Retrieved from https://www.astronomer.io/blog/airflow-in-action-apple/
2. Apple Machine Learning Research. (2024). *Introducing Apple Foundation Models*. Retrieved from https://machinelearning.apple.com/research/introducing-apple-foundation-models
3. Shrestha, R. (2024). *Big Data Analytics and Its Use by Apple*. Retrieved from https://www.slideshare.net/RakxitShrestha/big-data-analytics-and-its-use-by-applepptx
4. Apache Airflow Documentation. (2024). *Apache Airflow: Workflow Orchestration for Data Science*. Retrieved from https://airflow.apache.org/