Elevating Data Quality: Great Expectations and Airflow at PepsiCo

Russell Lamb
Introduction

WHO AM I?

Name
Russell Lamb

Role
Retail Data Engineering Lead in PepsiCo's eCommerce

Mission
Deliver retailer data to drive value for eCommerce

Based in
Me: New York - Team: Asia, Europe, USA

Favorite PepsiCo Products
SodaStream + Cool Ranch Doritos

Why do I love Airflow?
Flexible, scalable, extensible
PepsiCo: Global Presence & Powerful Portfolio

- More than $79 billion net revenue in 2021
- 23 billion-dollar brands
- More than 200 countries & territories
- Creating Growth & Value
Consumers Get our Products via 3 Main Channels

**Retail**
Consumers buy our products to consume at home. Many types of Retailers, offering a wide variety of brands and flavors.

**Foodservice**
Consumers buy our products outside of a retail location to consume on site. Usually our brands are exclusive at that site.

**Ecommerce**
Consumers buy our products via online merchants to have either delivered to their home or pick-up in store.
800+ active Production Airflow pipelines
- Self-hosted Airflow on managed Kubernetes
- Multi-cloud: AWS and Azure
- Dedicated data platform team in-house

50+ Data Sources
- Mostly customers, some 3rd party providers
- Multiple data types & formats

Multi-team Airflow instance
- Tag Airflow jobs by team
- Mono-repo: Shared code repository with several teams

Decentralized Development
- Data Vault Airflow plug-in
- Distributed ownership by domain
- Collaborative governance
- Shared data models
Data Quality – Challenges

- **Schema Changes**
  - Monitoring needed for consistency
  - Data is 2\textsuperscript{nd} and 3\textsuperscript{rd} party

- **Bad Data Hard to Remove**
  - Requires changes to production
  - Custom logic used once

- **Docs for Data**
  - Capturing the business context
  - What does good data look like?

- **Errors**
  - Missing Data
  - Duplicates
Our Solution

- **Great Expectations**
  - Version 0.13, July 2020
  - 300+ Expectation suites
  - Soft fail vs Hard fail

- **Custom Code**
  - Expectations Operator for Cloud Storage and Database
  - Action class for Slack and Rendering
  - Action for Data Docs Rendering

- **Why**
  - More notification options, e.g. multi-channel
  - Handling Authentication, especially on Azure

- **Many options “out of the box”**
  - GreatExpectationsOperator
  - Webhook for Slack
  - Use existing data docs
What is Great Expectations?

Great Expectations
• Python library
• Open Source

Components
• Web App: Data Docs
• Validation Engine
• Notifications via Webhook

Integrates with Airflow
• Validation step in pipeline
• Open Source Operator
• Custom Operator
Three-Tiered Approach to Quality

Inbound Data
Transforming Data
Warehoused Data

Great Expectations
Data Build Tool
Data Observability
Why Data Quality is Important

- Investing in data quality systems allows us to focus on building new capabilities
- Gartner: Bad data costs companies an average of $12.9 Million per year \(^1\)
- IBM: Bad data costs the US $3.1 trillion \(^2\)

Confidence in Data
- Sales decisions
- Marketing Spend
- Inventory replenishment

Save Time and Resources
- Logic of each must be documented
- Documentation should change with code, not in a separate document to maintain
- Good documentation allows engineers to build off the efforts of their peers

Variety of Providers, Sources, and Use Cases

---

\(^1\) Tech Target – data quality - [https://www.techtarget.com/searchdatamanagement/definition/data-quality](https://www.techtarget.com/searchdatamanagement/definition/data-quality)

Validation Lifecycle

- **Inbound Data**: Airflow senses new data
- **Define Expectation**
  - Rules data must follow
- **View Results**
  - Results in JSON or HTML format
- **Apply to Data**
  - e.g., Expect column values to be unique
- **Notify Data Owner**
  - Reach out to customer or vendor
- **Attempt Resolution**
  - For small issues like file name
Running in Production

Slack Alert notifies Data team of failed validation

Data analyst clicks link to Great Expectations “Data Docs” site

Data Docs contains readable results of the evaluation

Check the Airflow logs to verify the error

```
"result": {
  "observed_value": [
    "asin",
    "date_shipped",
    "fulfillment_center",
    "item_name",
    "merchant_brand_name",
    "shipped_cogs",
    "shipped_ops",
    "shipped_units"
  ],
  "details": {
    "mismatched": {
      "missing": [
        "program"
      ]
    }
  }
}
```
Benefits of Great Expectations

- Gives clarity to non-Subject matter experts
- Which columns can be null?
- What’s a valid value?

- Find the reason for failure quickly
- Embed Great Expectations “Data docs” links in Airflow log & Slack

- Stop pipelines before loading data
Questions?

Find me on LinkedIn
www.linkedin.com/in/russelllamb