Data Quality and Validation in ML Pipelines

Great Expectations, Deequ, and TensorFlow Data Validation

By: Sunil Kumar Mudusu

Why Data Quality Matters

Models Need Good Data

ML models need good data to learn and work well

Bad Data = Bad Predictions

Errors, wasted time, broken models

Can't Fix With Better Models

Even small issues in data can break your model

Common Data Problems

Missing Values

Empty fields or null values

Wrong Data Types

Numbers stored as text

Schema Drift

Structure changes over time

Outliers

Strange values breaking patterns

What is Data Validation?

Verify Data Quality

Check data is clean, complete, correct

Early Detection

Catch problems before model training

Automation

Can be automated in every ML pipeline

Meet the Tools

Deequ

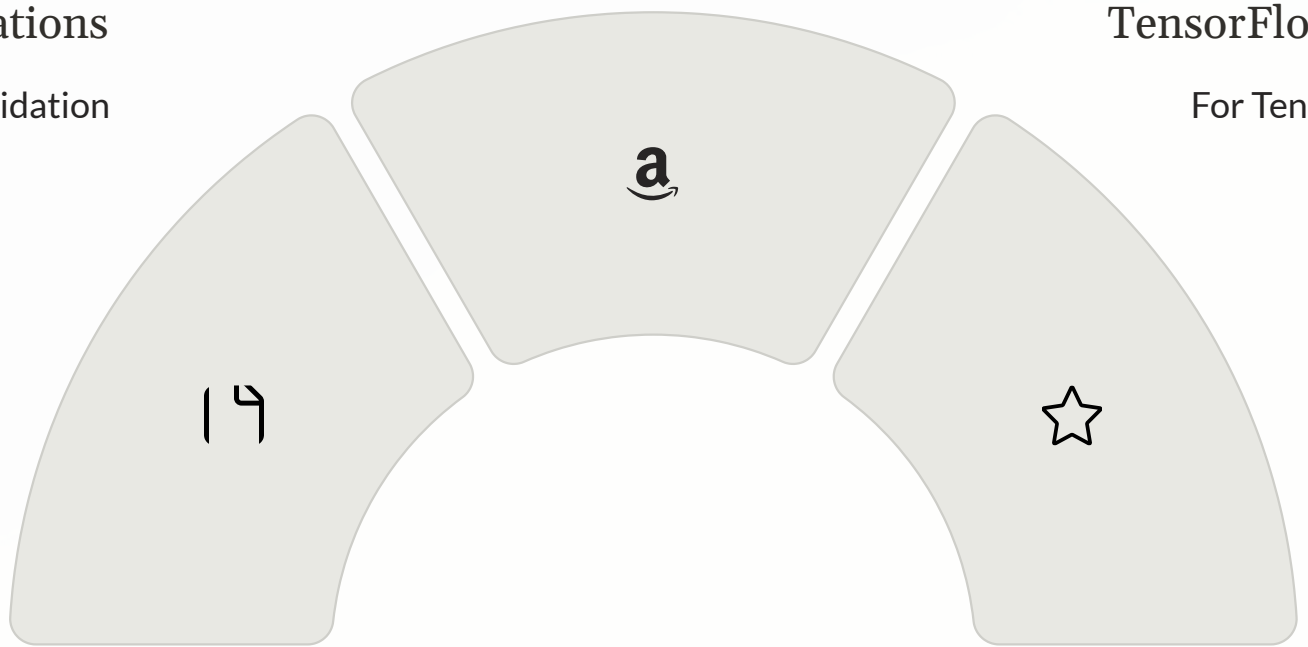
Spark-based for big data

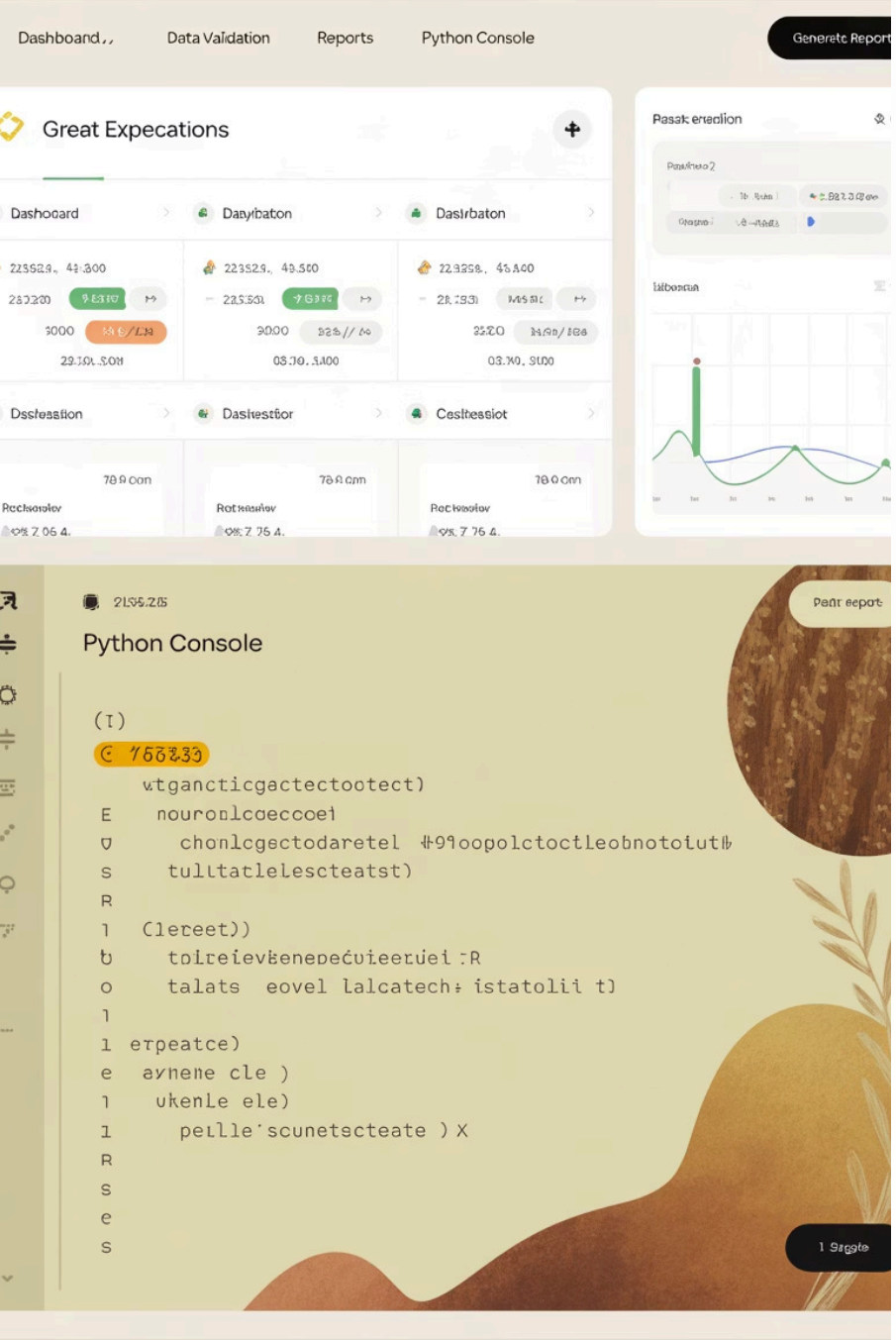
Great Expectations

Python-based validation

TensorFlow Data Validation

For TensorFlow pipelines





Great Expectations

Python-Based

Familiar for data scientists

Easy Rules

Simple data expectations

Clear Reports

Human-readable results

Deequ

Amazon-Built

Runs on Apache Spark

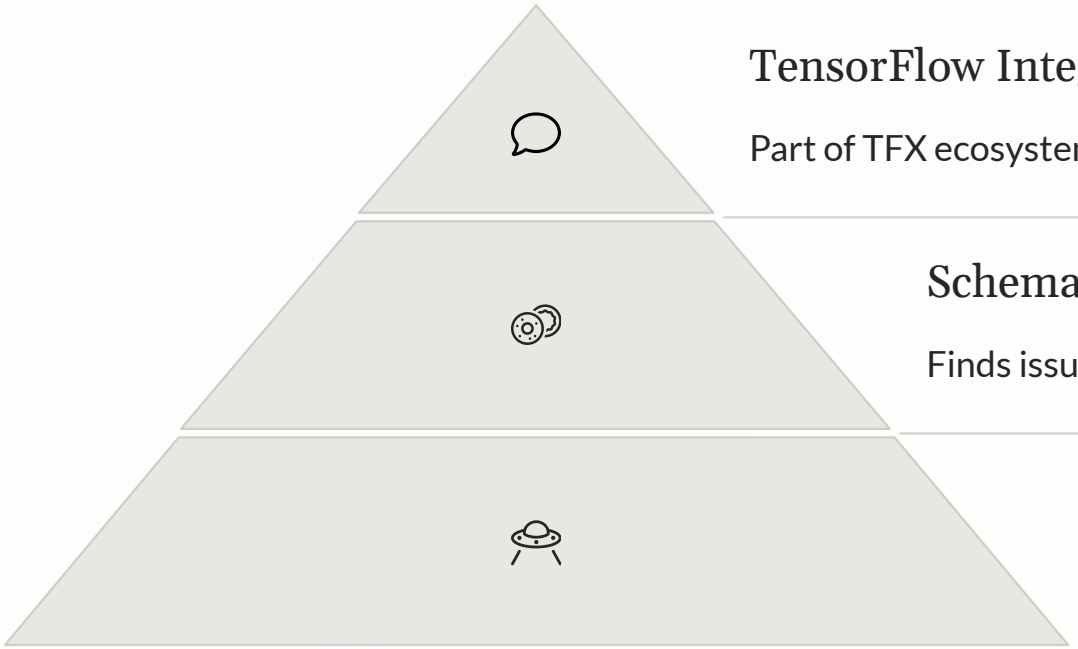
Big Data Ready

Handles large datasets efficiently

Quality Metrics

Tracks data quality over time

TensorFlow Data Validation



TensorFlow Integration

Part of TFX ecosystem

Schema Detection

Finds issues automatically

Apache Beam

Handles big data processing

Comparing the Tools

Tool	Language	Best Use
Great Expectations	Python	Easy validation, reporting
Deequ	Scala/Python	Big data, streaming
TFDV	Python	ML pipelines

Choosing the Right Tool

Assess Your Needs

Consider data size, tools, team skills

Match Tool to Workflow

Great Expectations: readable reports

Deequ: large-scale data

TFDV: TensorFlow pipelines

Implement and Iterate

Start simple, expand validation over time

Key Takeaways

Data Quality Critical

As important as model quality

Silent Problems

Data issues harm ML results

Multiple Options

Each tool has different strengths

Use Automated Tools

Check data before training

Conclusion

✓

Good Data

Leads to better models

🛠️

Right Tool

Choose for your workflow

Thank You