Thank you sponsors!











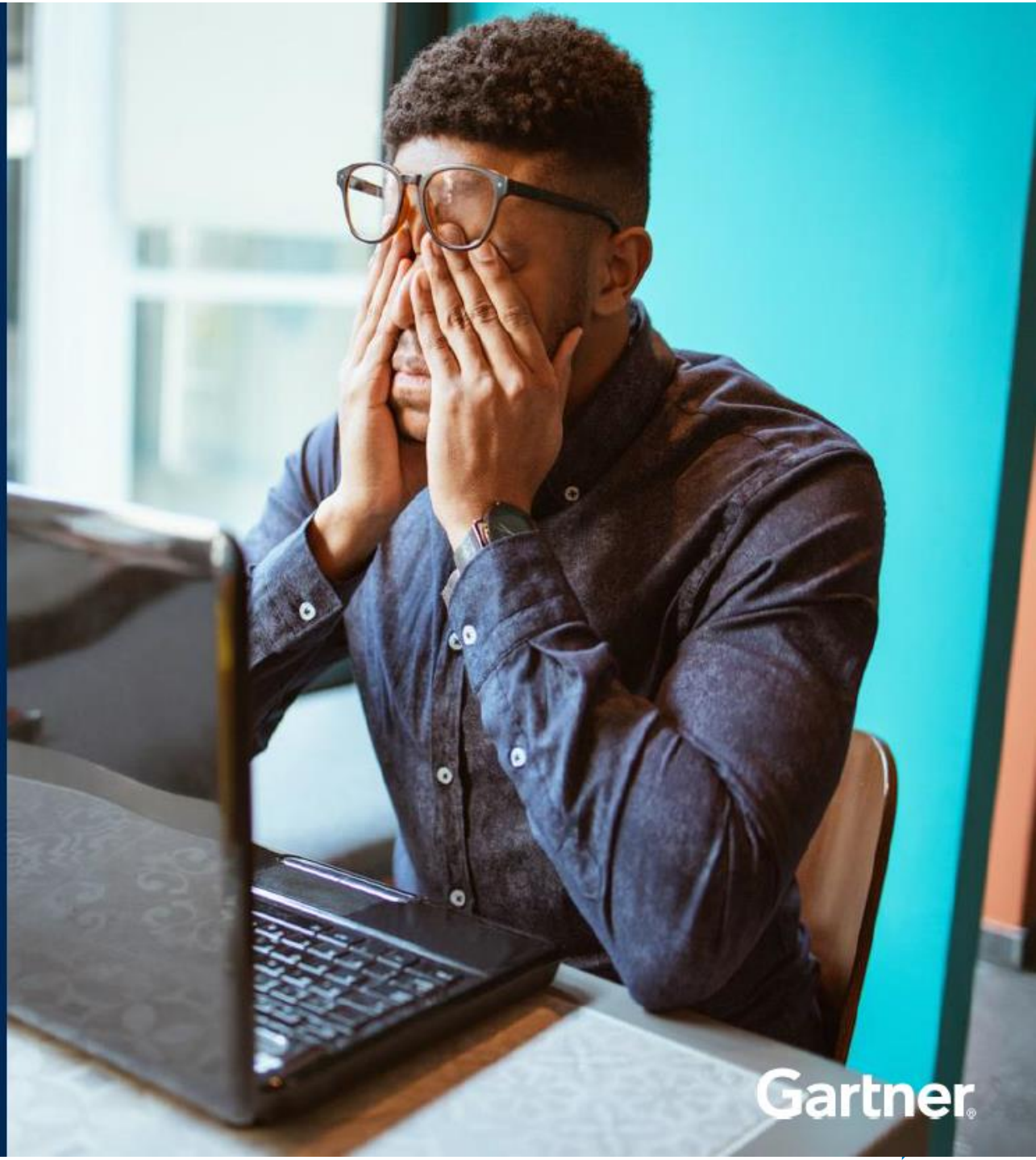




Only 7%

of data teams could resolve data issues before they impact users.

Source: Data Observability Core to Data Strategy for 92% of Leaders: CDO Magazine & Kensu Report, CDO Magazine



© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Gartner

Unseen Failures

Adventures in data pipeline testing

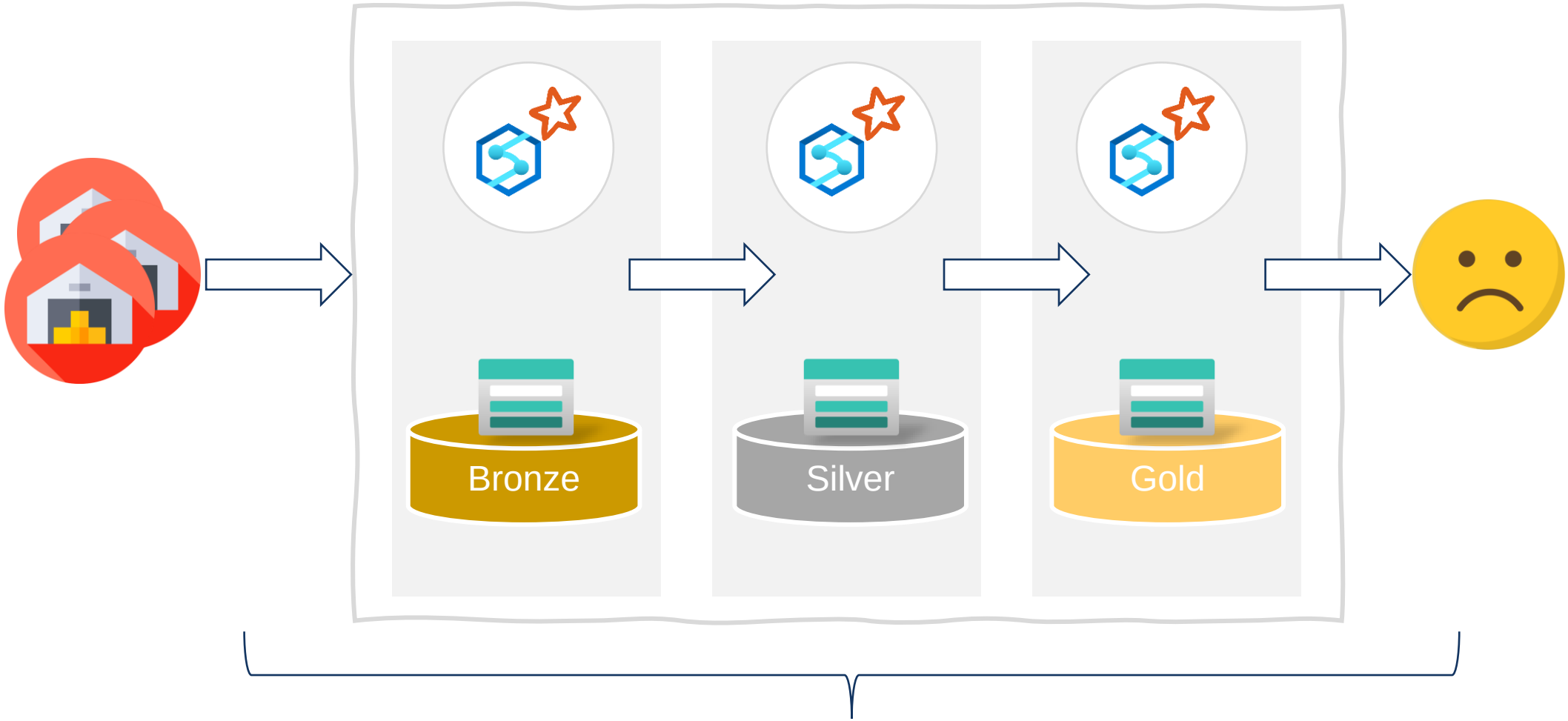


Mark Streutker

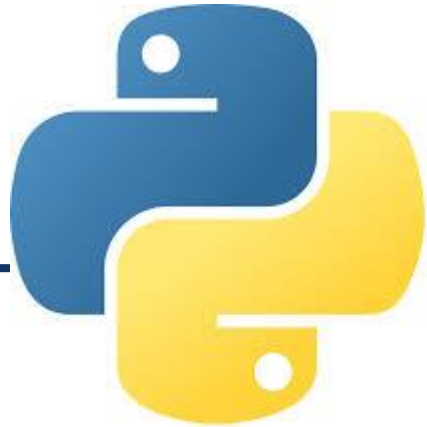
IT Consultant
Architect
Spreker
Chapter Lead Data & AI

infoSupport
Solid Innovator

Architecture



Technology







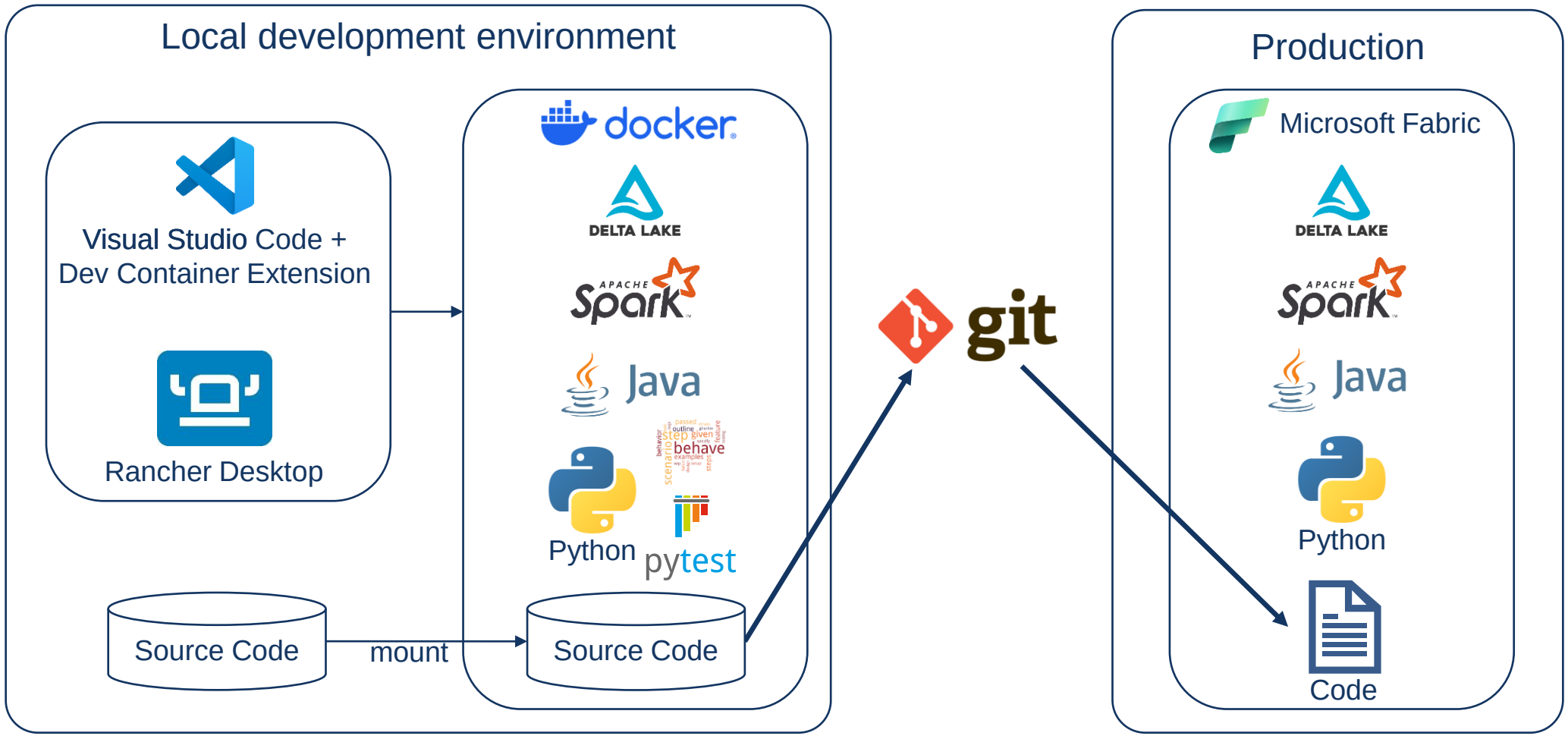
databricks

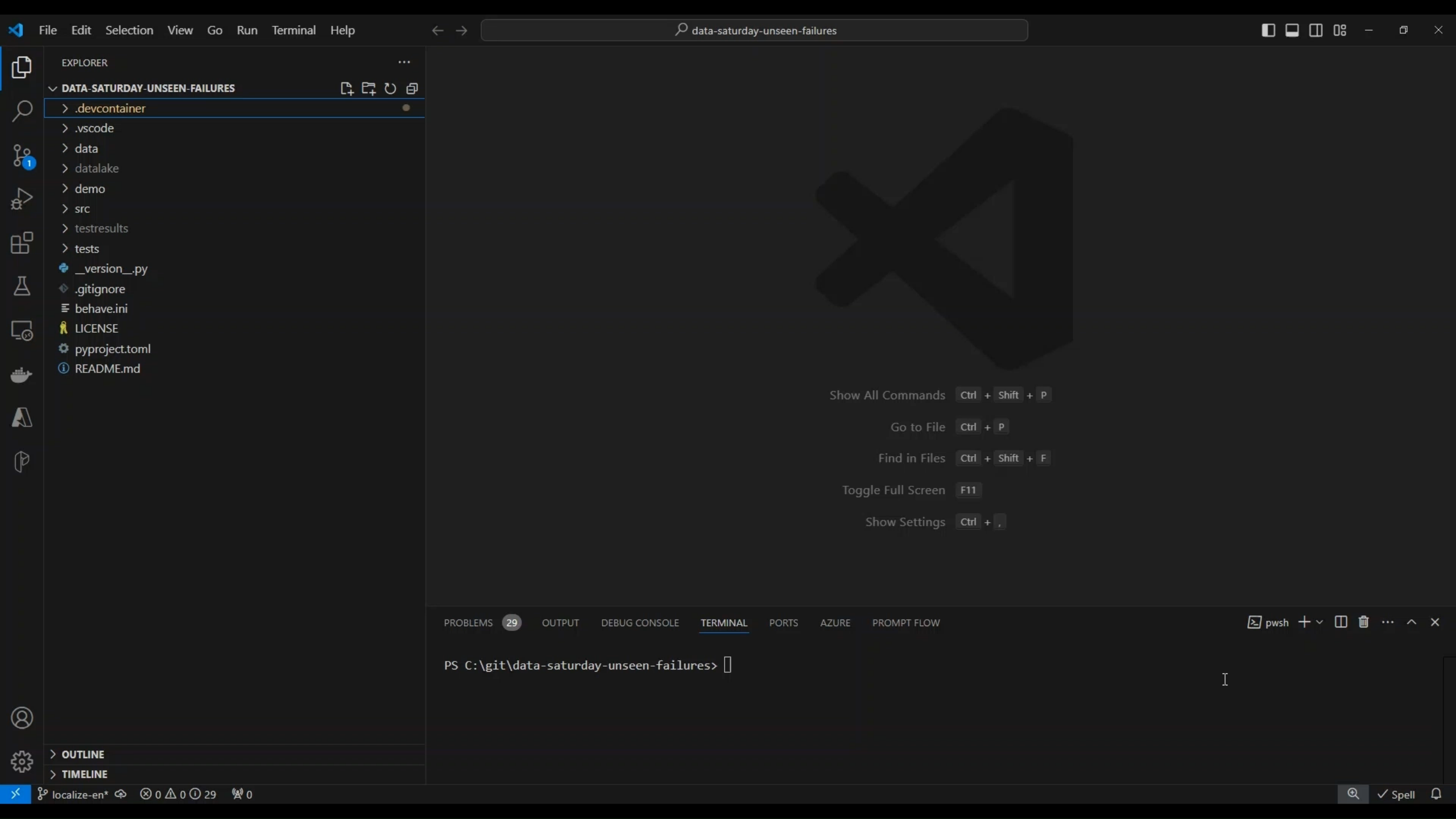


DELTA LAKE

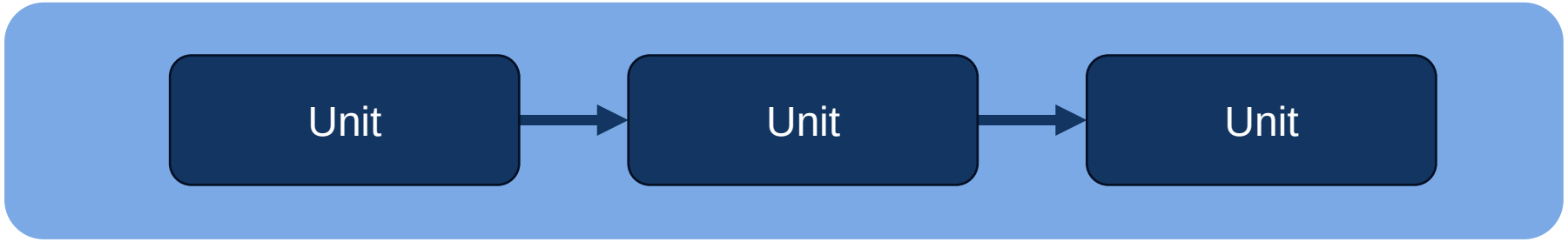


Solution


Local development environment
Visual Studio Code +
Dev Container Extension
Rancher Desktop
docker
DELTA LAKE
APACHE Spark
Java
behave
Python
pytest
Source Code
mount
Source Code
git
Production
Microsoft Fabric
DELTA LAKE
APACHE Spark
Java
Python
Code

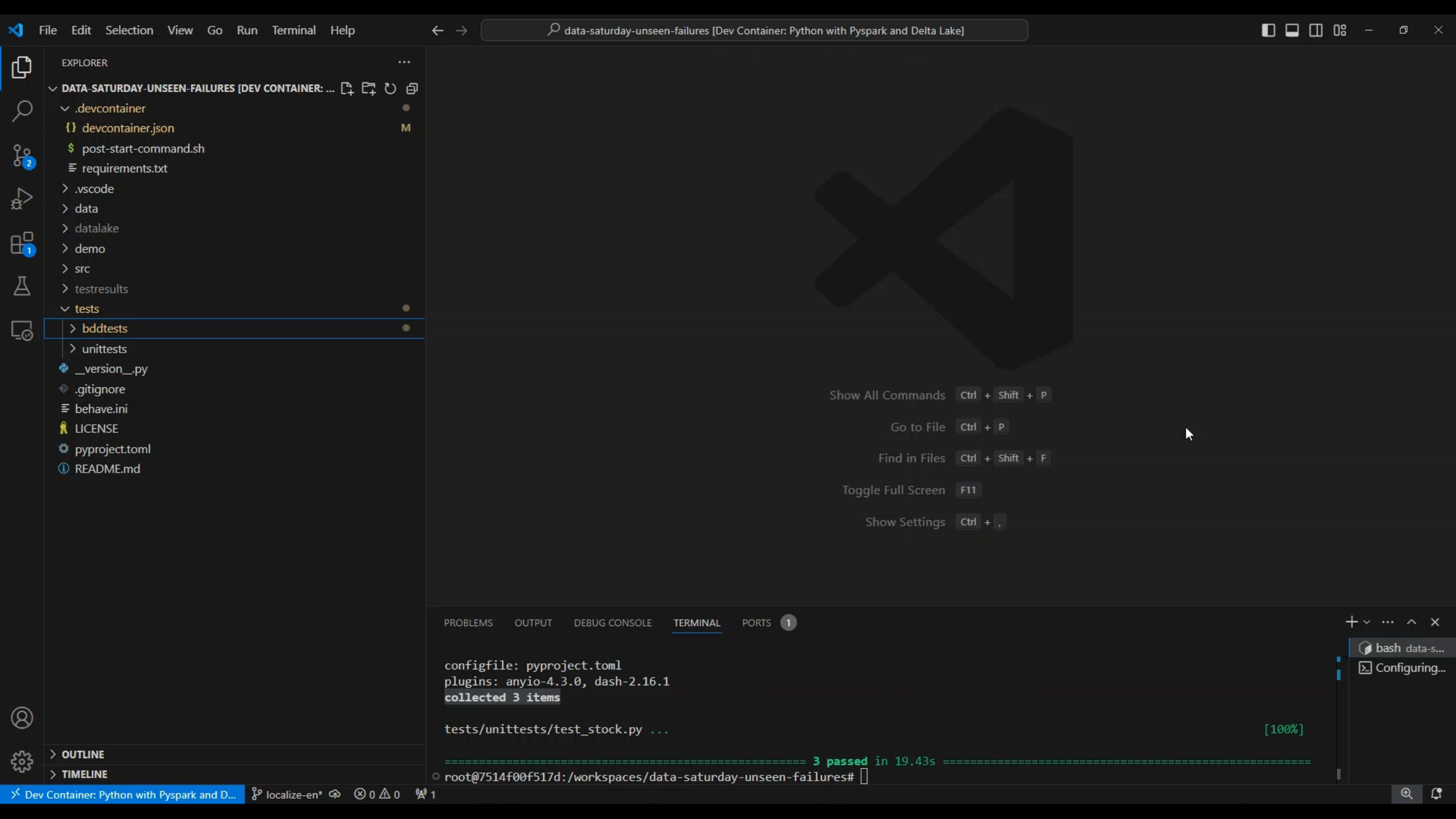

File Edit Selection View Go Run Terminal Help
data-saturday-unseen-failures
EXPLORER
DATA-SATURDAY-UNSEEN-FAILURES
.devcontainer
.vscode
data
datalake
demo
src
testresults
tests
__version__.py
.gitignore
behave.ini
LICENSE
pyproject.toml
README.md
Show All Commands Ctrl + Shift + P
Go to File Ctrl + P
Find in Files Ctrl + Shift + F
Toggle Full Screen F11
Show Settings Ctrl + ,
PROBLEMS 29
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
AZURE
PROMPT FLOW
pwsh
PS C:\git\data-saturday-unseen-failures>
OUTLINE
TIMELINE
localize-en*
0 0 29
0
Spell

BDD

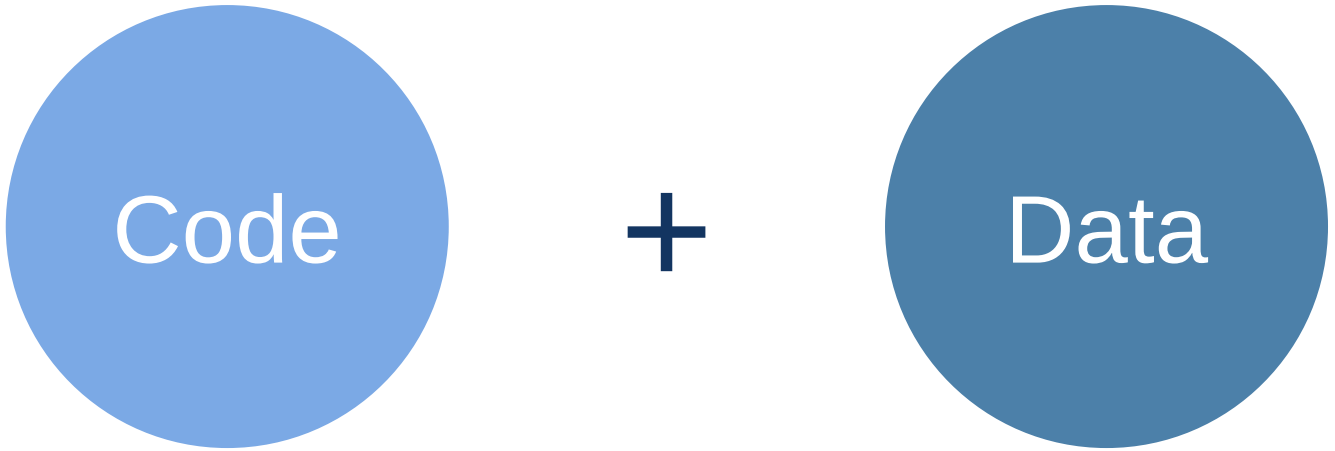


Behavior-driven development

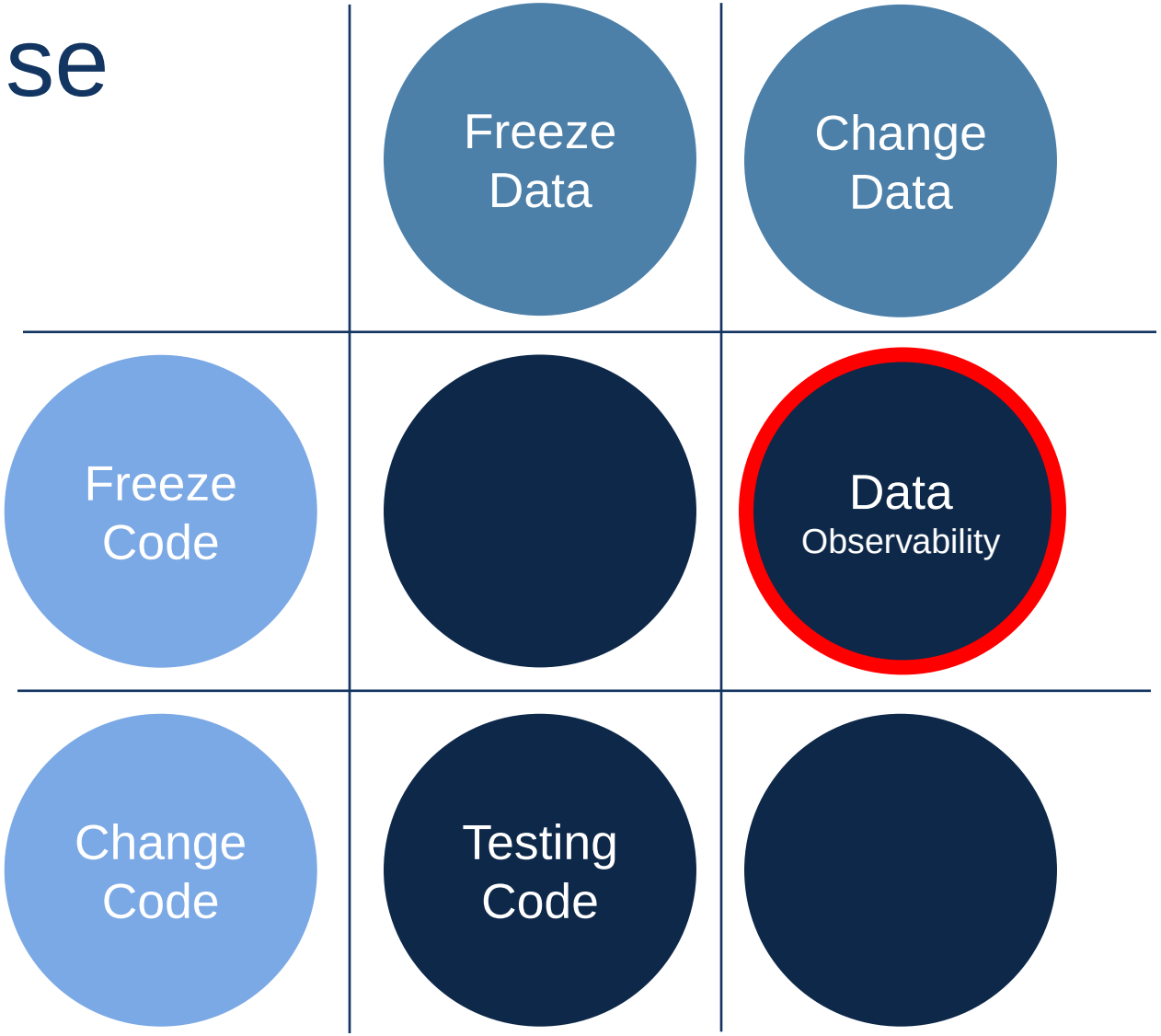
- Natural language
- Express behavior and expected outcomes


File Edit Selection View Go Run Terminal Help
data-saturday-unseen-failures [Dev Container: Python with Pyspark and Delta Lake]
EXPLORER
DATA-SATURDAY-UNSEEN-FAILURES [DEV CONTAINER: ...
.devcontainer
devcontainer.json M
post-start-command.sh
requirements.txt
.vscode
data
datalake
demo
src
testresults
tests
bddtests
unittests
__version__.py
.gitignore
behave.ini
LICENSE
pyproject.toml
README.md
Show All Commands Ctrl + Shift + P
Go to File Ctrl + P
Find in Files Ctrl + Shift + F
Toggle Full Screen F11
Show Settings Ctrl + ,
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 1
configfile: pyproject.toml
plugins: anyio-4.3.0, dash-2.16.1
collected 3 items
tests/unittests/test_stock.py ... [100%]
3 passed in 19.43s
root@7514f00f517d:/workspaces/data-saturday-unseen-failures#
bash data-s...
Configuring...
OUTLINE
TIMELINE
Dev Container: Python with Pyspark and D...
localize-en*

Root cause



Root cause


Freeze Data
Change Data
Freeze Code
Data Observability
Change Code
Testing Code

Data Observability



Data Observability

Is your data updated at the right time?



Data Observability

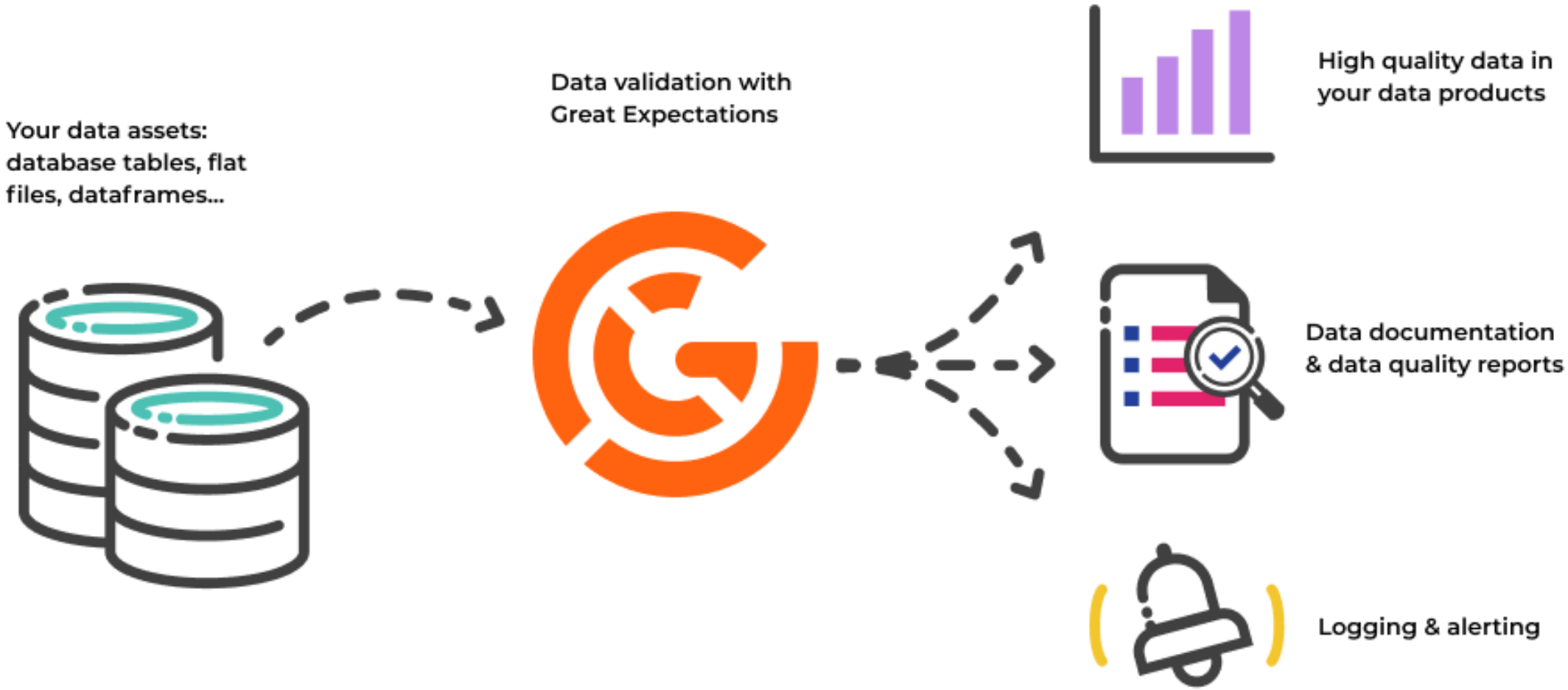
Is your data complete?

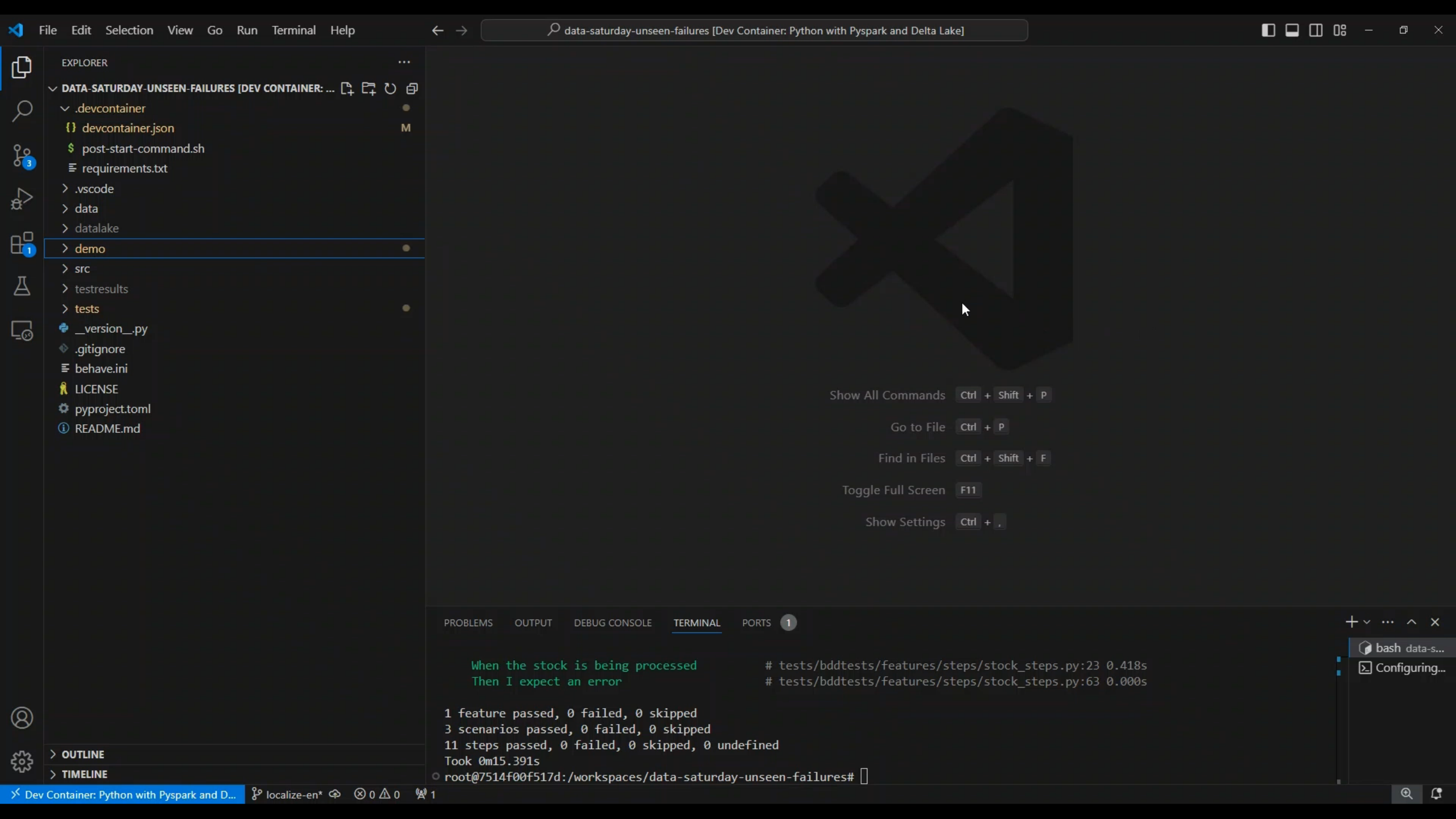


Great Expectations



How Great Expectations works




File Edit Selection View Go Run Terminal Help
data-saturday-unseen-failures [Dev Container: Python with Pyspark and Delta Lake]
EXPLORER
DATA-SATURDAY-UNSEEN-FAILURES [DEV CONTAINER: ...
.devcontainer
devcontainer.json M
post-start-command.sh
requirements.txt
.vscode
data
datalake
demo
src
testresults
tests
__version__.py
.gitignore
behave.ini
LICENSE
pyproject.toml
README.md
Show All Commands Ctrl + Shift + P
Go to File Ctrl + P
Find in Files Ctrl + Shift + F
Toggle Full Screen F11
Show Settings Ctrl + ,
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 1
When the stock is being processed # tests/bddtests/features/steps/stock_steps.py:23 0.418s
Then I expect an error # tests/bddtests/features/steps/stock_steps.py:63 0.000s
1 feature passed, 0 failed, 0 skipped
3 scenarios passed, 0 failed, 0 skipped
11 steps passed, 0 failed, 0 skipped, 0 undefined
Took 0m15.391s
root@7514f00f517d:/workspaces/data-saturday-unseen-failures#
bash data-s...
Configuring...
OUTLINE
TIMELINE
Dev Container: Python with Pyspark and D...
localize-en*
0 0
1

great expectations

Data Docs autogenerated using Great Expectations.

Actions

Show Walkthrough

Data Docs | my_data_docs_site

Validation Results | Expectation Suites

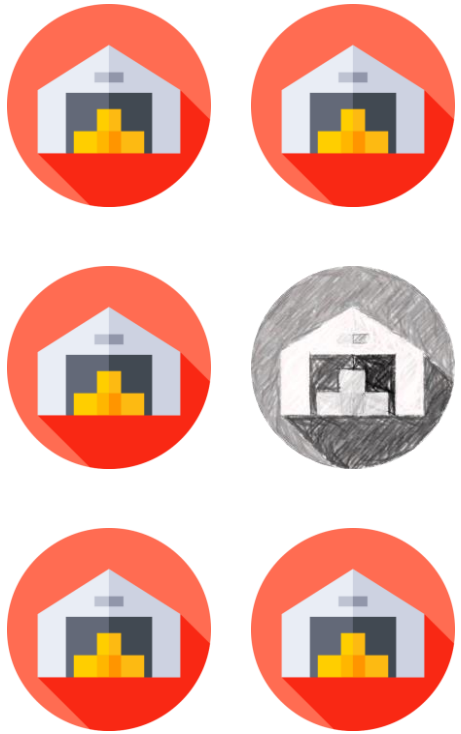
Search | Clear Filters

Status	Run Time	Run Name	Asset Name	Batch ID	Expectation Suite
✓	2024-09-28 15:01:29 UTC	test-warehouse_inventory_farmsum.csv	inventory	pandas-inventory	great-expectations-demo
✗	2024-09-28 15:01:28 UTC	test-warehouse_inventory_emmen.csv	inventory	pandas-inventory	great-expectations-demo
✓	2024-09-28 15:01:27 UTC	test-warehouse_inventory_delft.csv	inventory	pandas-inventory	great-expectations-demo
✓	2024-09-28 15:01:27 UTC	test-warehouse_inventory_culemborg.csv	inventory	pandas-inventory	great-expectations-demo
✓	2024-09-28 15:01:26 UTC	test-warehouse_inventory_breda.csv	inventory	pandas-inventory	great-expectations-demo
✓	2024-09-28 15:01:26 UTC	test-warehouse_inventory_amsterdam.csv	inventory	pandas-inventory	great-expectations-demo

Showing 1 to 6 of 6 rows

Stay current on everything GX with our newsletter Subscribe

Root cause





Key Takeaways

- Be aware of the different types of tests
- Start out small



Sourcecode

Find the demo at:

https://github.com/mstreutker/data-saturday-unseen-failures



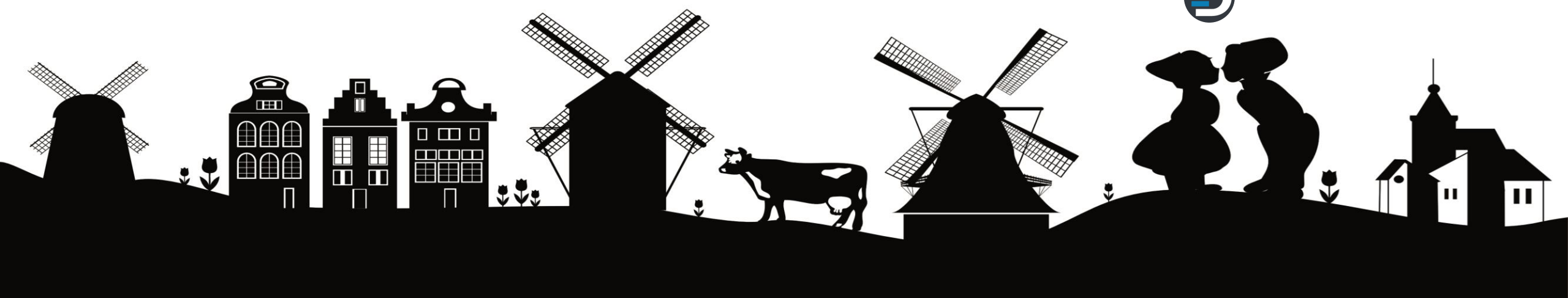
Rate Data Saturday Holland

SUPPORT CASPER

1 review = 1 €

Towards beating pancreatic cancer

Rate this session




InfoSupport
Solid Innovator