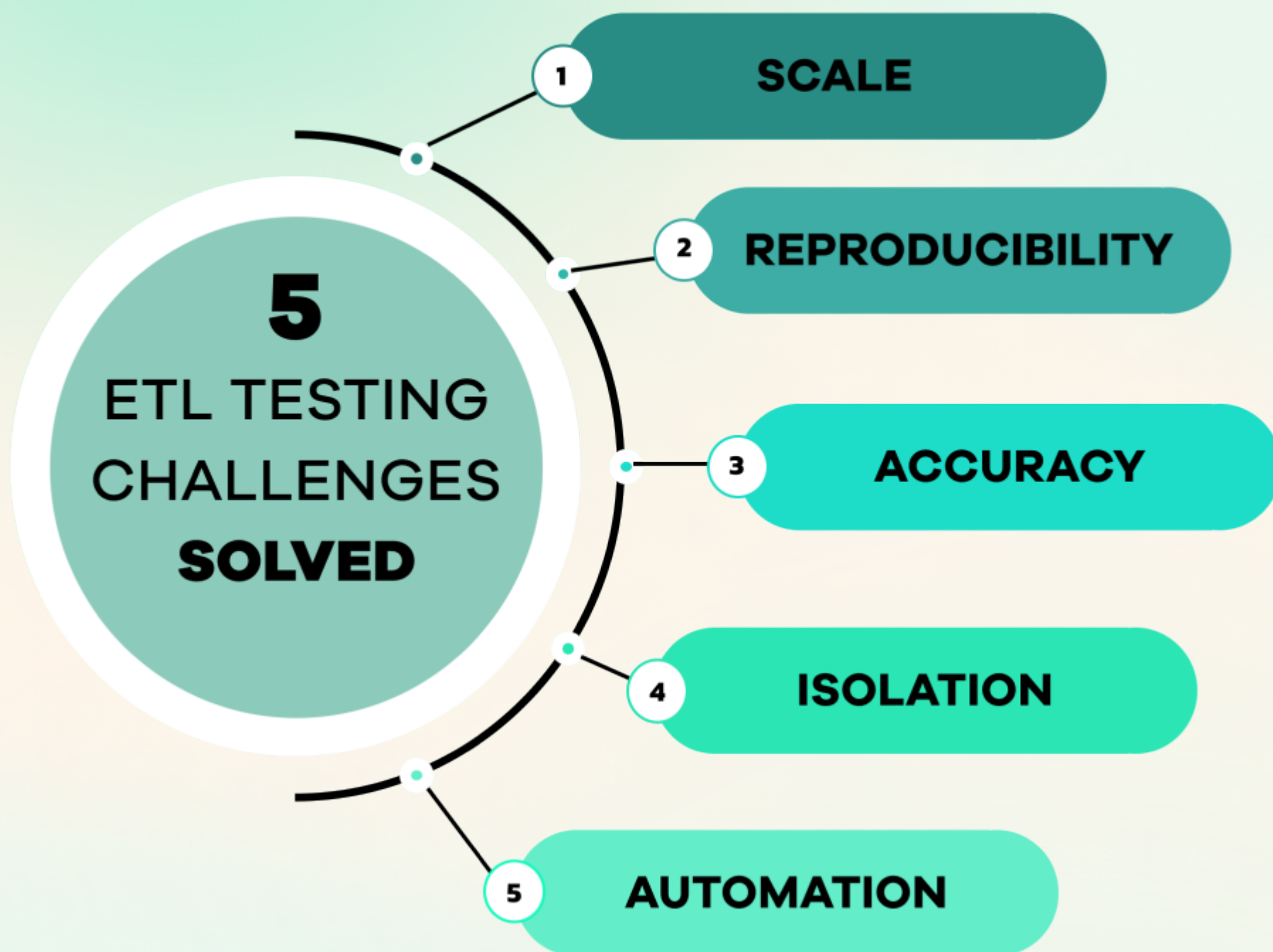
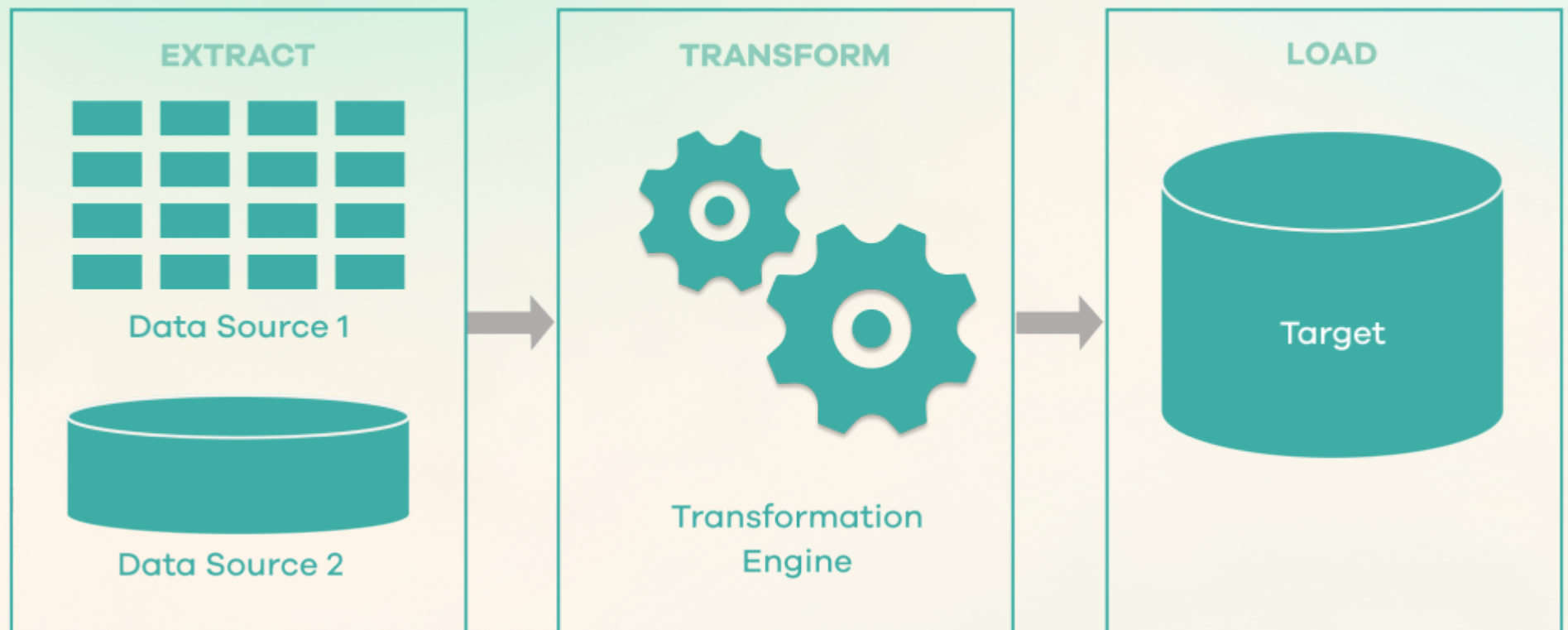


How to solve the most common ETL challenges with lakeFS



We'll explore the 5 most common challenges in ETL testing and then provide solutions on how to overcome these obstacles.



Source: <https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl>:

For the full article, visit:

<https://lakefs.io/blog/etl-testing>

SCALE

Replicating production data at scale is a time-consuming and expensive process.

How do you test against petabytes of data without copying it over and over again?



SCALE

lakeFS's data versioning takes advantage of metadata manipulation, making it easy to branch out of production as many times as you want.

Scale petabytes of data faster—without any excessive costs.



REPRODUCIBILITY

Data constantly changes but the output of an ETL pipeline shouldn't necessarily change when running the same data input.

How do you repeatedly test ETLs against the same data set?



REPRODUCIBILITY

Every new version of your ETL can be easily tested against historical commits with lakeFS.

Compare the output of different versions of your ETL against the same input with a repeatable process...



ACCURACY

The complexity and variety of data dramatically impacts ETL performance.

How do you test against production-like data?



ACCURACY

With the lakeFS branching, ETL testing is done against identical production data including all the complexities that come with it.

Get a snapshot of your real production environment without impacting ETL performance.



ISOLATION

Testing ETLs directly from your production object store, while doable, is risky.

How do you test complex ETLs without risking production data?



ISOLATION

Create fresh isolated test branches from your production data and remove the threat of risking production data.

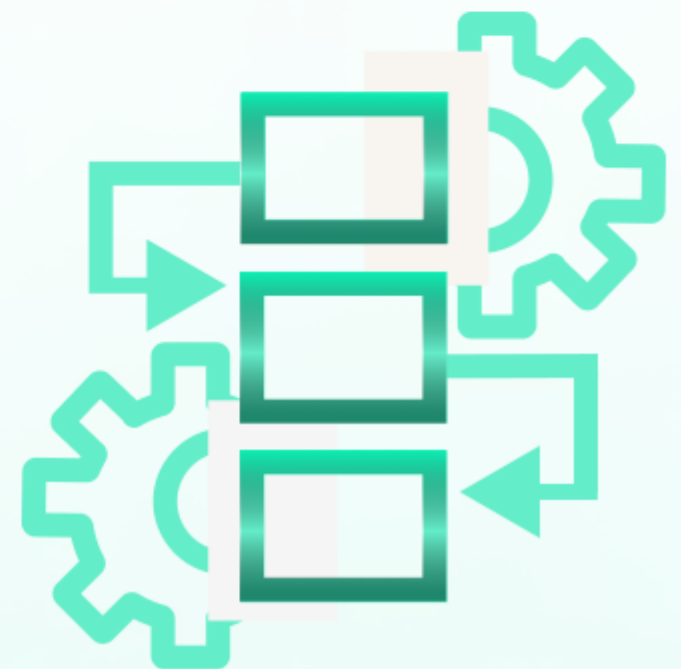
Start testing your new code against a fresh branch with peace of mind.



AUTOMATION

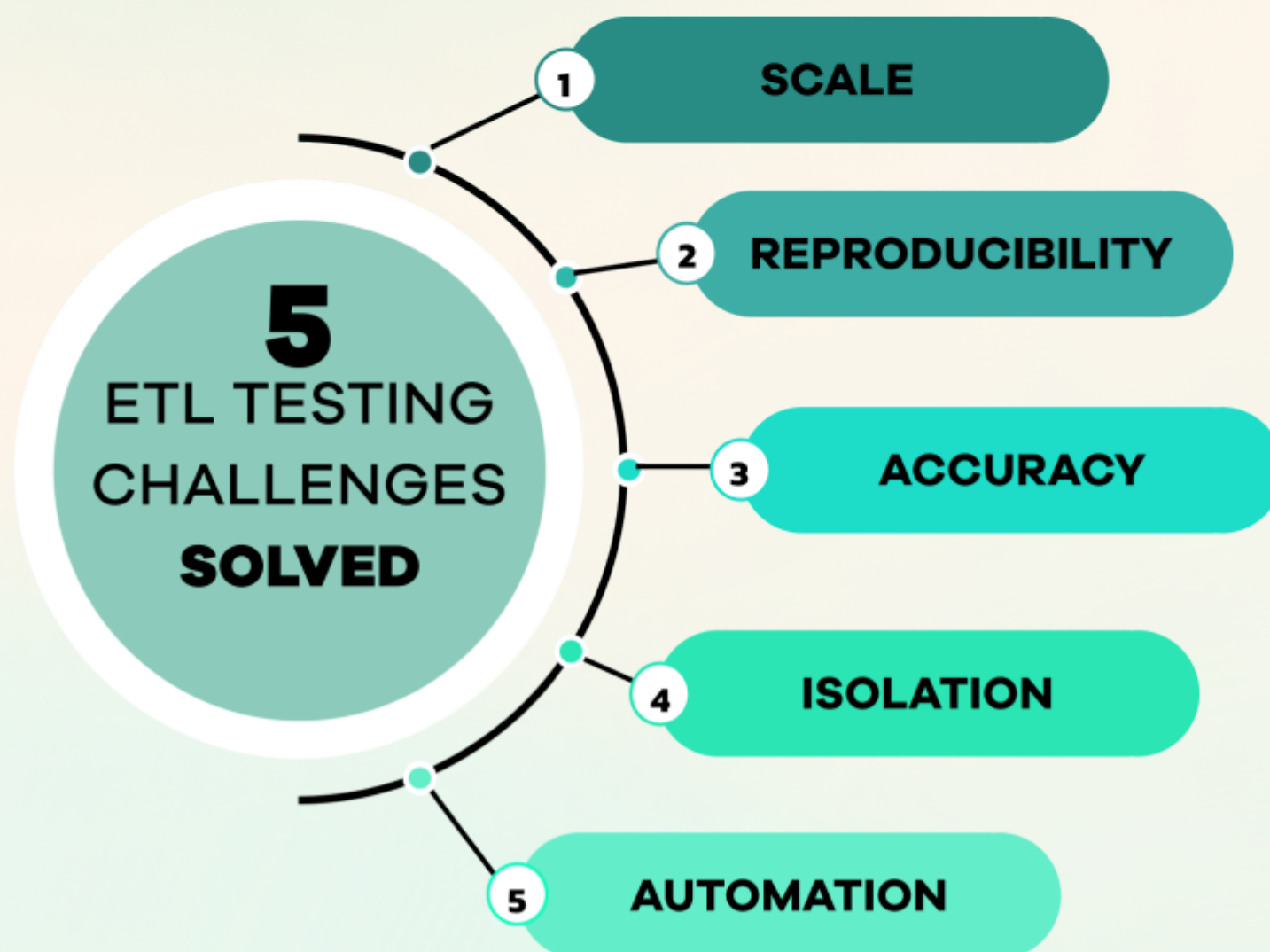
ETLs are long-running, multi-steps, expensive processes.

How do you automatically detect defects while testing ETLs—early?



ETL testing holds unique challenges beyond a “typical” software test case.

This article addressed the 5 most common challenges we’ve run across and then dove into how lakeFS can help you deal with these obstacles.



About lakeFS:

[lakeFS](#) is an open source and cloud data version control platform that provides data engineering best practices using git-like operations on data.

To learn more about lakeFS, check out our [playground environment](#) (no installation required) or talk to one of our [data experts](#).