Soda SQL Launch Blog

Let Soda SQL Do Your Data Testing!

Has it happened to you? We've seen it happen many times, maybe not to you, but to others we know.

What have we seen? We've seen too many inhouse attempts at data testing that just haven't worked out as well as people had hoped. Or what's worse, they haven't worked at all.

But instead of complaining about this, we rolled up our sleeves and decided to do something about it. So we created Soda SQL, a library for testing, profiling and monitoring the data in your data pipelines.

Silent Data Issues Can Wreak Havoc

In software, as in so many other areas, what you don't know *can hurt you*. At Soda, we refer to these unknown things as *silent data issues*. Even with data engineers in the front line protecting against them, silent data issues can wreak havoc on the data that is passed to your customers downstream.

The first line of defense is to check the data that enters your pipelines and the data that exits your pipelines. We call this *data testing*. Let's take a look at how Soda SQL lets you define data testing that detects issues before they become fledged problems.

Soda SQL Defends Against Silent Data Issues

Soda SQL works hand-in-glove with Data Engineering workflows. As an engineer, you get full control and visibility. You define how Soda SQL works by using industry standard YAML configuration files. These files can be checked into version control and let you control the tests that are executed and the metrics that are used to evaluate the results.

When new data enters your pipeline, for example, a Soda SQL scan will build queries that are based on the YAML configuration files. These queries are optimized to gather the most relevant information with regard to the data that is being tested. This information is used to compute metrics that you can use to determine if there are issues with the data. If there are issues, you can then stop the data pipeline.

OK, But How Does It Work?

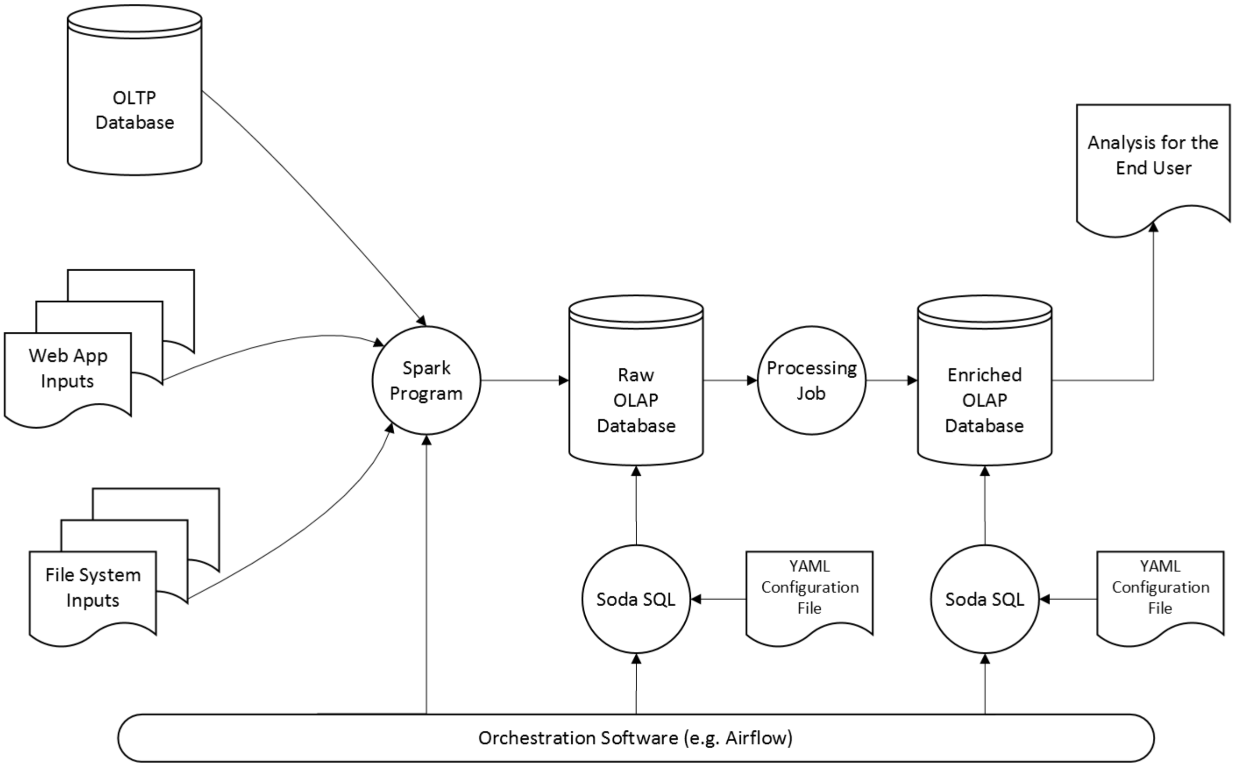
Maybe we should unpack that last section a bit. What exactly do we mean by data pipeline? When does Soda SQL run? How are the YAML configuration files used? How does anyone really benefit from data testing?

Let's start with a definition. A *data pipeline* is an operation where data is

- extracted from a database or
- transformed outside of a database or
- loaded into a database

In other words, a data pipeline is one of the constituent operations of an Extract Transform Load (ETL) or Extract Load Transform (ELT) process.

When does Soda SQL run on a data pipeline? To answer this question, let's begin by looking at the diagram below:



Data Testing Workflow with Soda SQL

On the left, you can see the data from different sources that exists in an organization. Some data from Online Transaction Processing (OLTP) is stored in a relational database system, other stored data may be taken from web application inputs and still other data may be stored directly in a file system.

Periodically data from these different sources is aggregated and loaded into a raw Online Analytical Processing (OLAP) database, typically via a Spark program. This is a load data pipeline. It is on this pipeline that Soda SQL is first run, in order to catch the silent data issues that can cause so many problems.

The YAML configuration file, of course, defines how the Soda SQL job runs on the data pipeline. This file needs to be customized for every customer installation, since each organization has unique database schemas and unique data requirements.

When data is used for analysis, Soda SQL can also be run on the extraction data pipeline. A different YAML configuration file defines how this Soda SQL job will run. This file needs to take into account the schema of the enriched OLAP database. This may differ from the schema of the raw OLAP database.

So how does all this data testing benefit anyone? Maybe the biggest challenge in any organization is to analyze data and then make decisions based on what you discover. When the data has been tested, any analysis will be more accurate. This gives organizations the power to make better decisions.

The bottom line is that Soda SQL makes it possible to improve your bottom line!

So Why Call It Soda SQL?

As our name suggests, we went all in on SQL. Unashamed. Without any reservations. After the hype of NoSQL (which should have been called NoTransaction BTW) there is a clear trend back towards SQL across the full spectrum of data stacks. The warehouse is back, though in a completely different form. dbt is probably the best example of this trend and frankly, they have been a big source of inspiration for our approach. What's more, even data lakes are being equipped with SQL engines.

One other advantage of an SQL approach is that it lets you leave your data in place. You don't need to load or move your data around to test and monitor it. Soda SQL can simply be used by itself embedded in your pipeline for data testing.

Soda SQL and the Soda Monitoring Platform

We have talked about Soda SQL and how it lets you conduct data testing. Now let's talk about the Soda Monitoring Platform. We believe that we and our customers are all better off if we combine our data testing efforts with our monitoring expertise.

Maybe the simplest way to explain the relation between the Soda SQL library and the Soda Monitoring Platform is by analogy. If a car is made of an engine and a chassis, the Soda Monitoring Platform is the car itself and the Soda SQL library is the engine.

We're currently building a free cloud service that will store your metrics over time and enable anomaly detection, all as an extension of Soda SQL. Subscribe to our newsletter to stay up to date on these exciting developments!

But why wait for the newsletter? To learn more about Soda SQL, head over to
soda and check it out for yourself.