In an [earlier post](https://kjmackey.wordpress.com/2015/04/14/lakes-reservoirs-and-muffins/) I wrote about data lakes vs. data reservoirs. An effort to change the thinking by changing the words.

Continuing  with the theme of applying different metaphors, this post is about the idea of a data refinery. The metaphor is not new. It's already "out there."

Earlier this year I put together a data strategy for a company.

They have a fire-hose of data coming at them from multiple sources. They need to deliver to their customers valid and useful information siphoned off from the fire-hose. They also have a requirement, in a more lenient time frame, to deliver enriched datasets to those same customers.

Finally they need to use the enriched data, in aggregate form, to make better decisions about their own business.

Differing needs, differing levels of enrichment. Wash, rinse, repeat.

The focus of this entry is not the siphoning off of interesting data in near real time. That's its own entry for some other time.

This is about the dataset enrichment.

So you let your fire-hose pour into your data reservoir. Then what?

Traditional approaches run jobs across the dataset to extract information from it and ship said information to your customers. They're almost ETL jobs, just without the E part. They perform some transformations on the data, pull the enriched set out, and "load" it to the waiting customer.

In addition, analysts run some form of wide-ranging queries (however they might be formulated/encoded) to extract interesting patterns, answers to hypotheses.

My belief is that this is old thinking. The lake (sorry reservoir) is seen as a source and operated on to produce packaged results.

Newer thinking regards the systems around such a reservoir as forming a data refinery, and the processes applied to elements in the reservoir as data refinement. Continuous data refinement.

The enriched data remains in the reservoir and contributes to the next level of enrichment. Extracts are performed whenever a customer requires a dataset. The customer gets whatever the current state of enrichment provides. Analysts work with the enriched datasets or on aggregated sets of the enriched data.

The enrichment process is continuous. A framework needs to be put in place to support this, and logging should be performed to ensure repeatability. The framework needs to be extensible, because the enrichment process itself must be capable of enrichment over time. In another post I'll lay out what are the elements that belong in such a framework.

The refinery is where the company's data truly lives. It's where the information to support the company's business may be found.