Distributed Data Management

dataset.io

Dataset IO Suite 100. 222 South Church St, Charlotte, NC **T** 704-298-1234 **E** hello@dataset.io www.dataset.io

Dataset IO was founded in 2012 to provide data management products and consulting to support the future state of analytics in the enterprise.

The new economics of utility computing and open source data platforms will change the way enterprises deliver services and handle the ever increasing complexity, speed and volume of data.

Reach us at:

Dataset IO LLC
222 South Church Street
Charlotte, NC 28202

T: +1 704 298 1234
hello@dataset.io

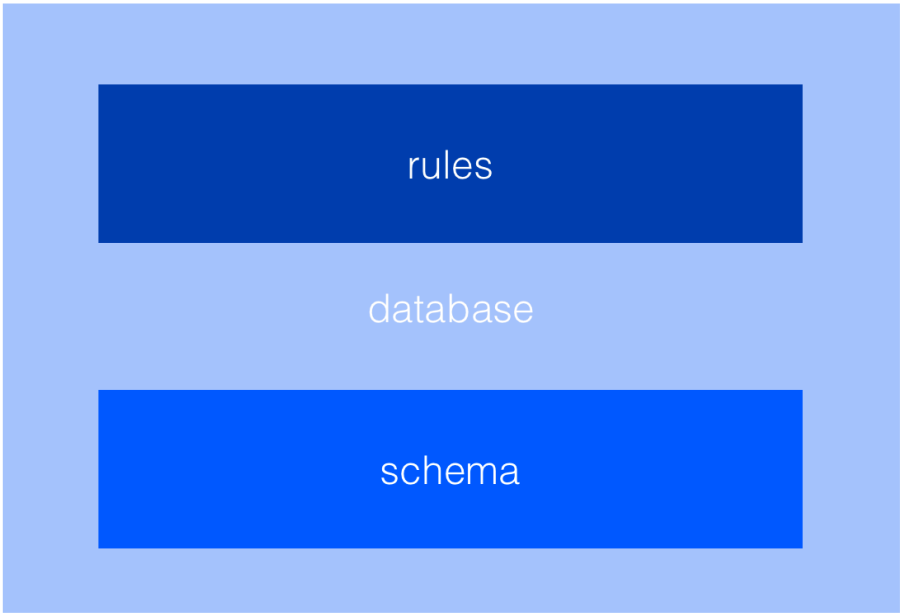
http://www.dataset.io

Data Management

A Brief History

The last thirty years can be seen as a generation based on the ubiquity of the relational database. Since the end of the 70's, the rise of the Oracle and the RDBMS marketplace has been the cornerstone of enterprise data management and application development.

The beauty of the database was that so much of the needs of data management, from data dictionary to validation rules, were taken care of in the platform. The model was simple and effective.

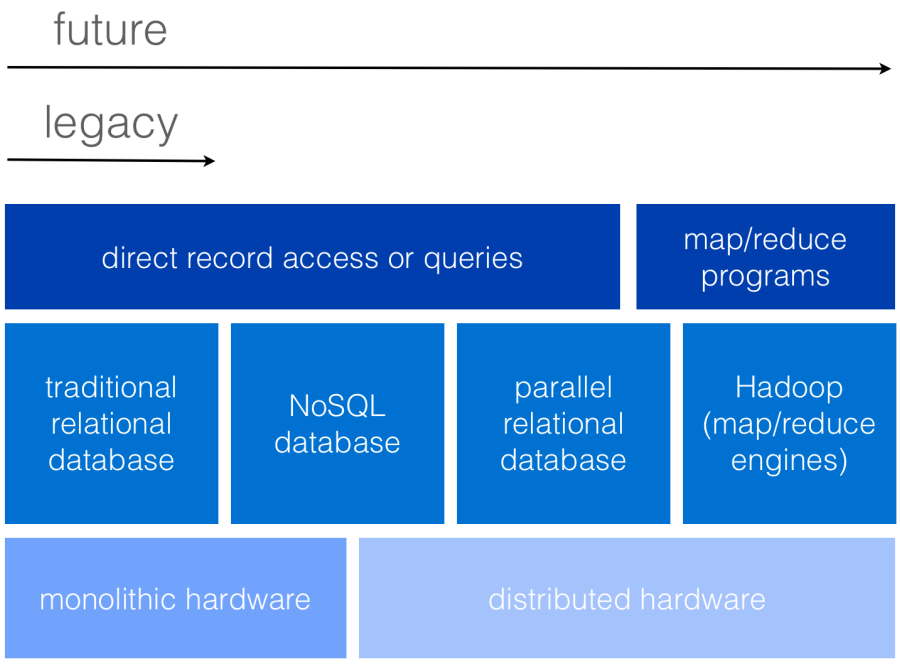


Where we started to have problems was that we started to see silo's of data appearing throughout the enterprise and we needed to build views that crossed those silos. Typically, that meant building a new RDBMS and using ETL tools to move the data from the silos to the central store - or Data Warehouse. The projects were often complex, very expensive, and the structure of the data and the mapping tended to be done in a programatic sense. ETL tooling was basically just a simple tool for pulling data, transforming it and reloading it.

The Data Warehouse was able to provide much of the reporting that a company needed - but typically this was still operational in nature and the analytics and insight that people wanted remained often just beyond the warehouse's design.

A New Landscape

Thirty years after Oracle was founded, in 2007, the technology world started to look beyond the confines of the relational database. Two new types of technology started to enter the conversation. The first was a set of data storage systems commonly referred to as NoSQL (due to the lack of SQL support) and the second was a new type of data processing pattern pioneered at Google called Map/Reduce and made available to all through a technology called Hadoop.



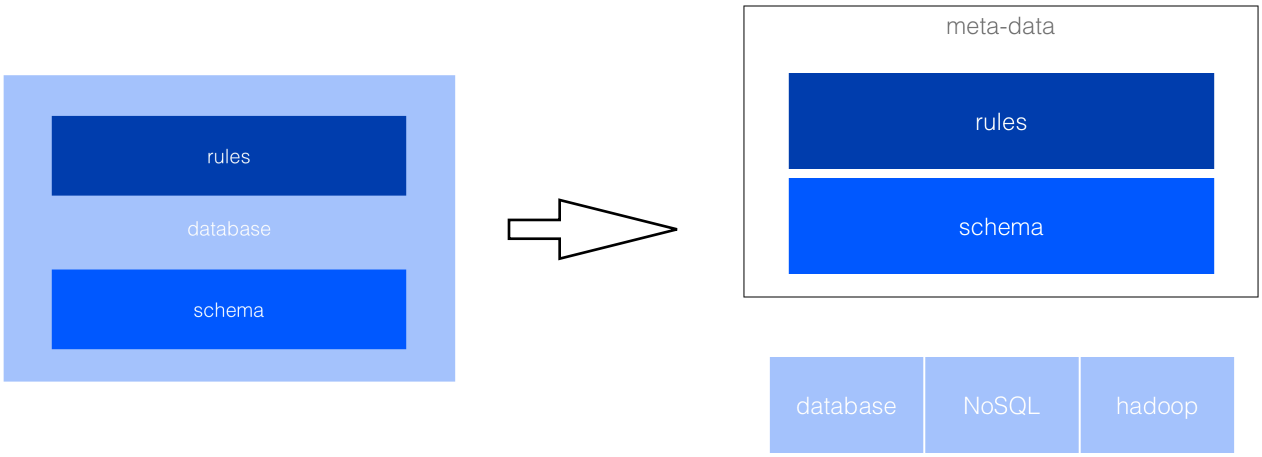
These new technologies suddenly enabled whole new types of data processing work and opened the door to new capabilities, such as:

- Full text search, allowing for huge amounts of information to be indexes outside of a database and searched with fuzzy logic
- The building of network graphs, showing the deep relationships between people, entities and more. New traversal queries could quickly operate on billions of relationships
- Document based databases allows for “soft schema” storage of data where new attributes and structures can be quickly loaded
- Big Data storage of hundreds of terabytes of data on commodity computing could be analyzed and queried at price points unimaginable with RDBMS solutions

Distributed Data Management

This new data technology landscape has meant that many of the traditional approaches to building a single universal data solution in your enterprise is now not practical. Your business will want to push for more powerful analytics and better insight and it is no longer a single data solution that will solve all your problems. In this world you are no longer aggregating data together somewhere - or building a data warehouse to the be basis of all your decision support solutions.

This means that the data management tools and capabilities that were the basis of your RDBMS world need to be pulled from it and exposed in a new way - a way that crosses the technologies available.



The rules and schema that your RDBMS provided now needs to be recreated in a way that can be shared across the technologies. You will need to be able to extract it from your existing data investments and then project that data structure into your new solutions.

But also the last thirty years have seen the field of data management advance and expand, and now it is more than just a schema and validation rules. Today, data management extends into understanding the transformations that your enterprise's data go through, the lineage of your data, the schedules on which data enters and moves through your world and more.

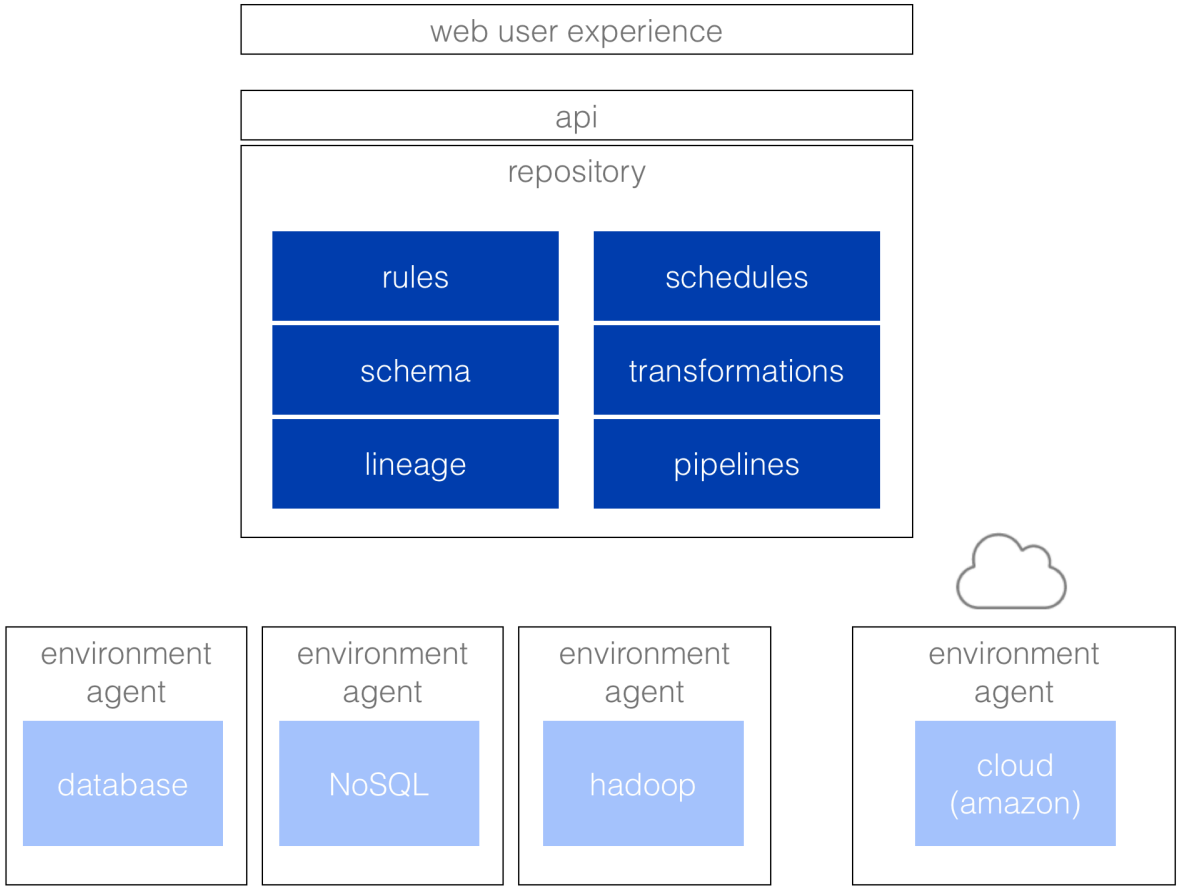
For analytics to be successful you need control over your data

This richness means that you are able to defend and trust the data and this is the most important aspect of data management and analytics. A data-driven enterprise is going to leverage a range of technologies and tools to be competitive - and the analytics space is evolving quickly. For analytics to be successful you need control over your data, from all its sources, and for this you will need distributed data management.

Dataset IO Platform

Our data management platform is designed based on a decade of experience building data solutions. The aim is to provide a simple and central way to pull together, manage, monitor and move data.

The heart of the architecture is a meta-data repository that enables you to capture schedules, data structures, data stores, transformations and more. This meta-data can be automatically harvested from files or other databases and then global data dictionaries can be built based on multiple sources.



Interactions with the data stores and the operation of all data flows are done through agents. An agent can be run by any machine, and also agents can be grouped together. While they use the repository server as a central store for all meta-data, they are independently able to pull and push data between each other. Meaning that when loading or moving data, from external files or from one store to another, agents can operate independently of the repository and thus remove potential bottlenecks.

Another powerful feature is the ability of the agents to be run in cloud environments, such as Amazon Web Services, where they can be deployed on-demand and leverage utility services (like Dynamo for data storage, S3 for files, or Elastic Map Reduce for on-demand Hadoop processing).

The architecture of our platform means that you can:

- Deploy agents near your existing data stores and use them to inspect the data schemas
- Use agents with schedules to source files from external sites (like FTP) and then push them into databases or maintain file stores
- Agents can be used to schedule the movings of data from one store (as an RDBMS) to another (Hadoop) - and is able to do it using the structures in the repository meaning often almost instant mapping
- Perform a cross-reference when you have multiple sources of data to combine data, agents are also able to run this against different types of architecture (from multi-threaded to Hadoop jobs)

The potential of being able to apply distributed data management across your environment to enable you to build solutions for data flows to analytics is a game changer for many enterprises.

With the Dataset IO platform in place you are able to:

- Quickly inspect each of your data stores and build a global data dictionary
- Use a web UI to build mapping across the discovered data structures
- Define and run cross references across your sources
- Schedule data pipelines to move data in or out of your enterprise
- Quickly push data into your Hadoop cluster for large scale processing
- Leverage the cloud seamlessly to do the heavy lifting of data cross-references and mapping

Today the technology offerings to push your data-driven enterprise forward are just starting to surface. Now more than ever is the time to start understanding, managing and leveraging the data buried deep in your world.