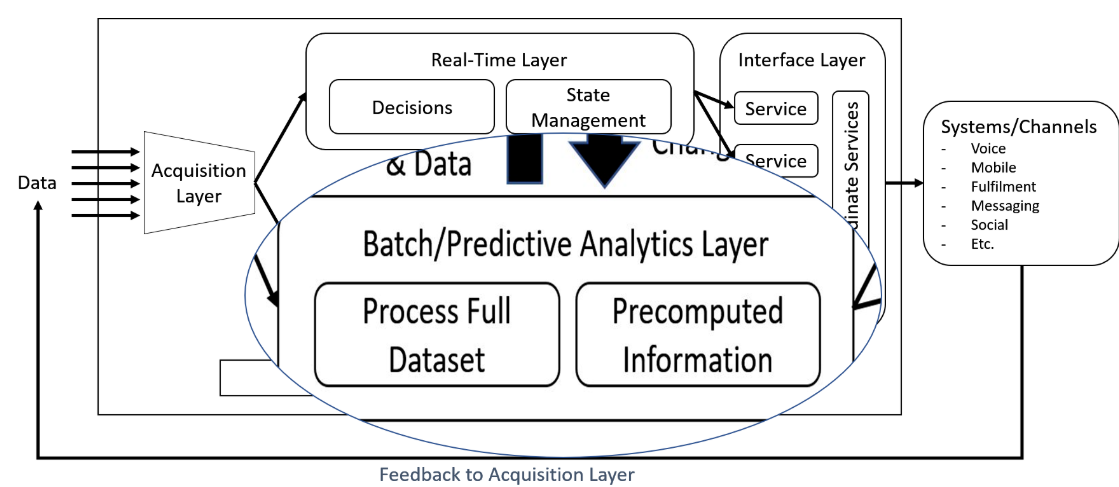
**Ab Initio Lambda - Batch Layer**

In this third article we follow on from the [Ab Initio Lambda Architecture](https://www.linkedin.com/pulse/ab-initio-lambda-overview-chris-day-/) and the [Acquisition Layer](https://www.linkedin.com/pulse/ab-initio-lambda-acquisition-layer-chris-day-/).

## Batch Layer Recap

The Batch Layer's focus is on serving business intelligence activities, and the majority of us will find comfort in this layer. The customers want a massive number of observations to make informed business decisions.



## Batch Layer in detail

The Batch Layer plays into Ab Initio’s core strengths; Processing data at scale

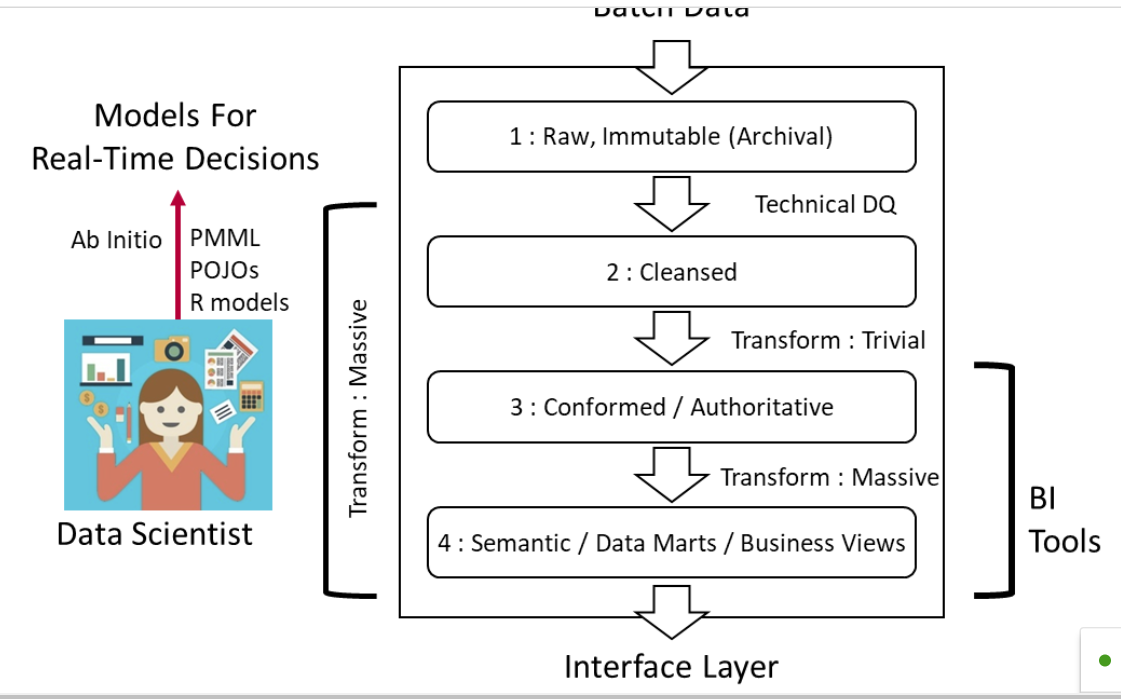
* store the raw, immutable and constantly growing data
* process or precompute the batch views for our customers

Nathan addresses this simply in his Lambda Architecture as

batch view = function(alldata)

## How does this look?

I would suggest this is akin to Extract, Validate, Transform, and Load (EVTL), something we are all familiar with.



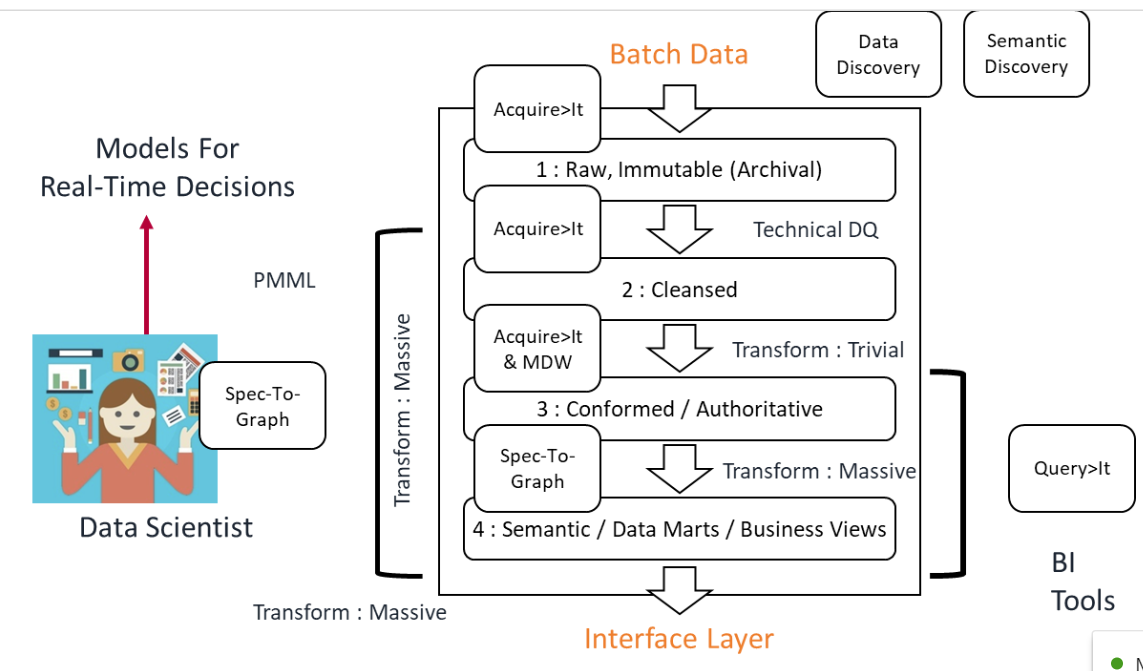
However, we have a seen a change in the batch layer due to the volume and veracity of incoming data, and the data analysts have evolved into data scientists, leveraging predictive analytics to support the digital transformation.

The challenges are numerous;

* A handful of data feeds now becomes thousands.
* Sporadic, negotiable change now becomes constant, enforced change
* Prescribed data extraction becomes data wrangling
* Data custodianship is currently regulated by law to protect our customers
* The lineage of both data and metadata to evidence trust and compliance
* And last but not least, finding and keeping the data scientists

Ab Initio addresses, I would say as always, one facet we can never have enough of, time. This is all down to the out-of-the-box parallelism at scale.

## What does this look like in our Ab Initio world?



With an end-to-end set of products and frameworks built from the ground up, you can give the business more time to make decisions to keep your information strategy ahead of the competition.

My bullet point observations from the community are 70% less calendar time to populate the data lake, 80% less time to fill the semantic layer and 75% fewer resources to complete the work.

Of course, the automated metadata management for Governance is still there.

### From Raw to Cleansed to Authoritative to Consumable

The components within the Ab Initio Lambda Batch Layer deal with obtaining the data in its raw, immutable format (which we archive), cleansing according to business rules, turning into conformed and authoritative information, then onwards to publish the semantic, Data marts and Business Views.

As one of the targets within the Acquisition Layer, we use Ab Initio's Acquire>It extensible framework to deliver both the raw, immutable data along with the cleansed data. Recall that Acquire>It’s focus is on providing simplicity for development and scalable at execution. To that end, it takes a single source, applies business rules and writes to many heterogeneous targets as you like (within our Lambda architecture that means the Batch Layer).

Along the way you’ll want to be initially applying some technical data quality rules, leveraging the power of the Data Quality Environment and using the Data Quality Assessment Framework. The outcome will be conformed, authoritative and trust data.

The final massive data transformation of multiple sources to multiple targets is built using Spec-To-Graph (This is an Express>It framework that automagically creates a graph). Ab Initio provides another timesaving benefit that aids your data wrangling.

## Companions to the Batch Layer

You’ll notice that the illustration contains many companion products to the Batch Layer.

We have Semantic, and Data Discovery to support the understanding of the data, we have Query>It to provide uniformity in extraction and Spec-to-Graph to keep the Data Scientists fed with as much data as possible.

Let’s jump into what these add to make the Batch Layer complete.

## Semantic and Data Discovery

### Data Discovery

Data Discovery delivers data profiles for the exploration of data before using Data Quality Assessment to compute detailed data quality measurements. (DQA was previously known as DQE, and now DQE is DQA + Data Discovery).

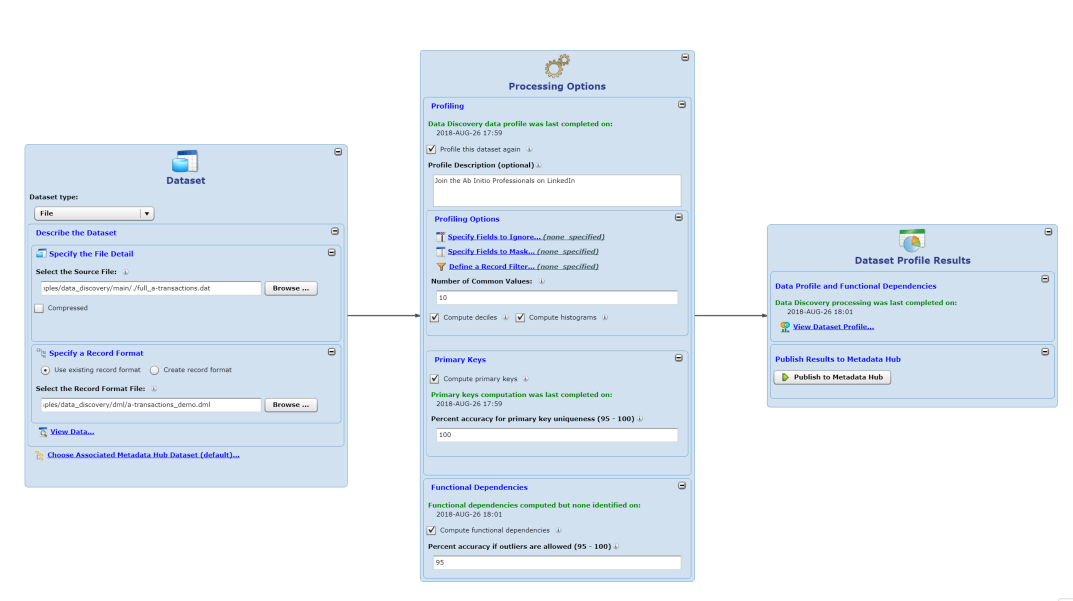
We begin by identifying each dataset in an application configuration and creating a data profile for an individual dataset. The profile for a particular dataset contains the following information:

* Statistics for the entire dataset
* Statistics for individual dataset fields
* Summary information about the diversity, validity, and completeness of values in the dataset fields
* Functional dependencies between dataset fields
* Primary keys in the dataset
* Masked data fields, so that the profile does not reveal personally identifiable information (PII)

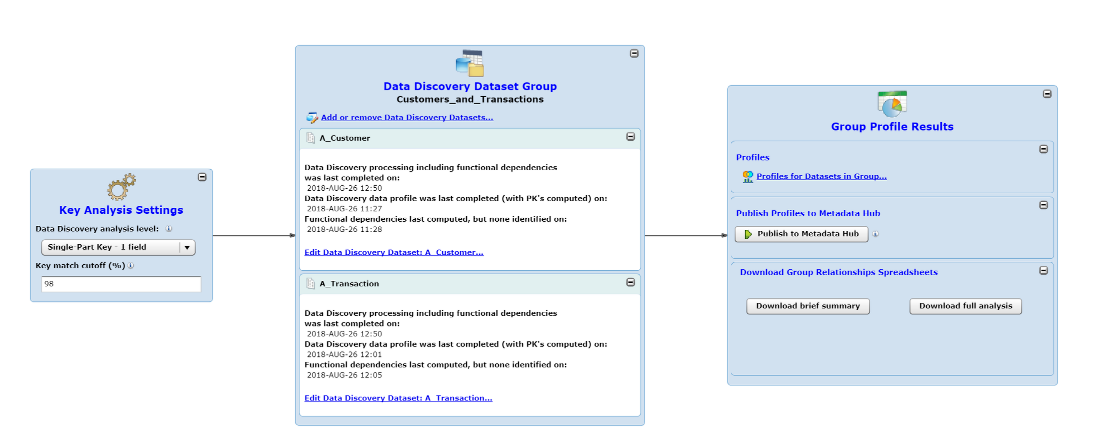
After you create individual datasets and profiles, you can create a dataset group. In a dataset group, you can create enhanced profiles that contain the following information about data relationships:

* Analysis of cross-field relationships between datasets
* Foreign key relationships between datasets

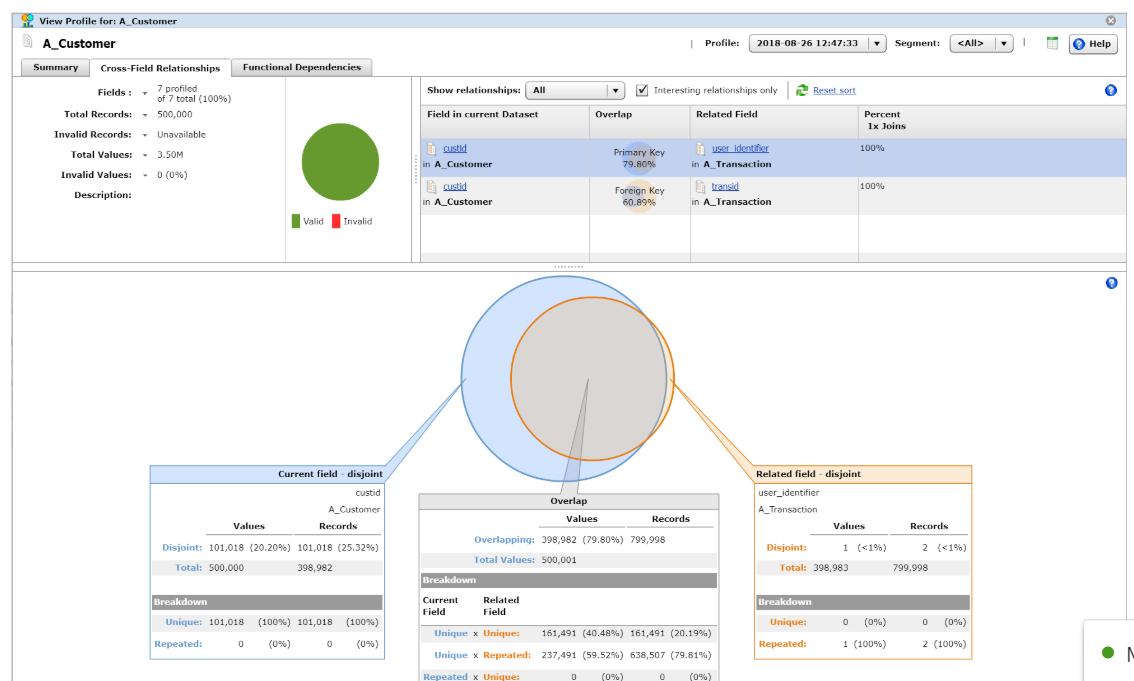
In this picture, we show the configuration of a dataset for discovery;



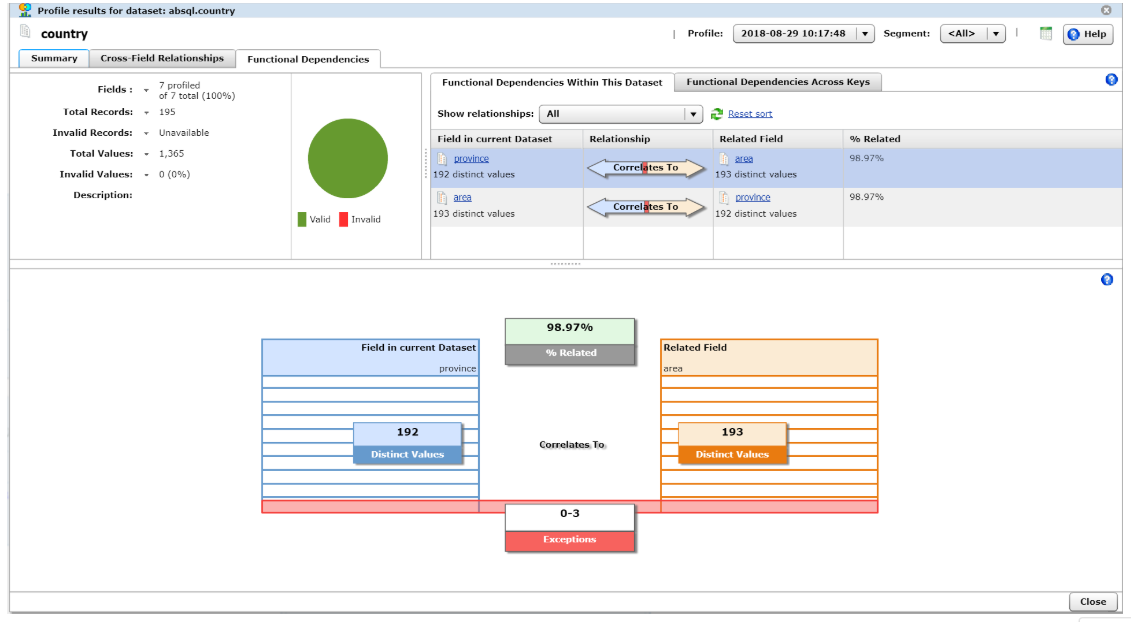
And then we can add multiple datasets into a Data Discovery Dataset Group;



Enabling us to show such insights as cross-field relationships



and functional dependencies used to find correlations in your data



### Semantic Discovery

Data is useless without meaning. Without a grip on the semantics, there is not a lot you can do with your data.

Semantic Discovery is all about finding those complex relationships between your data assets and helping attest they are relevant. The relevance is typically a score.

Metadata is a precursor to semantics, and this feeds an ontology-based approach to discovery. The Metadata>Hub’s Data Dictionary is a sensible starting point as it is transparent to the underlying technical syntax, communication protocols and formats used to describe the data. The next step would be to start enhancing the Data Dictionary with relationships between authoritative Domains, Data Elements, Data Sets and Logical Models.

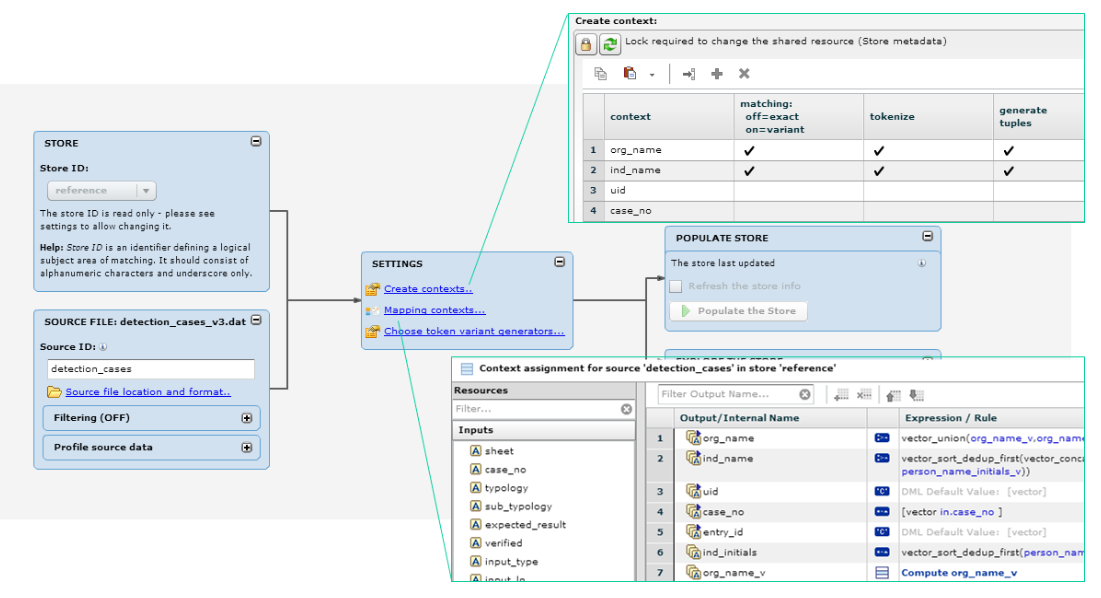
Once your ontology has momentum, the Semantic discovery process can produce a more consistent level of trust and relevance. The score is there to help make an informed decision, that’s all.

Ab Initio’s Semantic Discovery is a tunable, extensible black-box which helps you score or group records. It has 3 modes of execution;

* A Static comparison within a dataset or across datasets.
* Batch comparison of a new set of data against an existing master data set
* Real-time comparison of a new data item against a current master set using API’s.

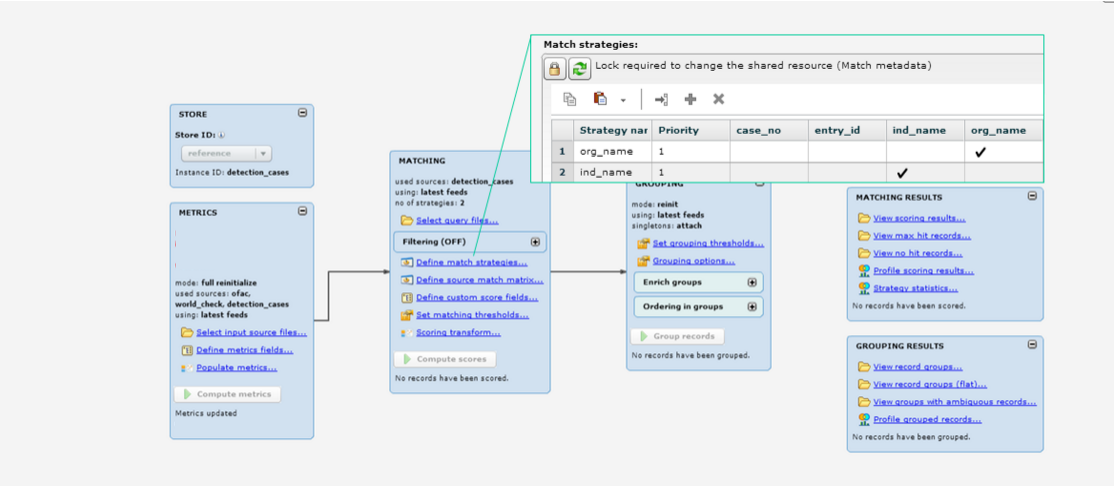
The scoring rules are built using the Business Rule Environment (BRE)

The Semantic Discovery process combines 2 steps; first, we define the dataset along with contexts;



Context is a logical grouping of fields, with associated strategies. A field may belong to more than one context. This is useful when comparing sources with different sets of fields containing similar information.

In the second step, we define the various matching strategies, that can be combined, to produce the results to help make an informed decision.



### Sensible, inclusive Data Provisioning with Query>It

With all the data you want to provide a scalable, robust, easy and secure way to deliver to your consumers, right?

Ab Initio’s Query>It is a high-performance, scalable SQL engine that can federate data from almost any structured or semi-structured data source. It allows you to query, move, and store data as well as to create and maintain indexed compressed flat files (ICFFs), Ab Initio's low-cost, low-latency mechanism for storing and accessing large volumes of archival data.

The multitude of sources Query>It can support include, but are not limited to;

* Tables in relational databases (such as Oracle, Teradata, and DB2)
* Flat files
* Files in a Hadoop Distributed File System (HDFS)
* Indexed Compressed Flat Files (ICFFs)
* Web services
* Any data source that Ab Initio can read

Query>It translates SQL into a Graph for execution and you can also save the generated graph too.

Another quick win, enabling consistent and straightforward access from R, Python, Power BI, Tableau and any other tooling that can talk SQL.

## Data Storage on the Batch Layer

Nathan writes a whole chapter on Data Storage within the Batch Layer. With Ab Initio, we don’t have to.

Ab Initio software was designed to be a parallel data processing framework, to work in a distributed environment, from the get-go.

Parallel, scalable distributed filesystems are there out of the box with the Co>Operating System through the Multifile System (MFS). Of course, you might want to put the Co>Operating System Multifile System on top of HDFS to benefit from data replication (HDMFS).

### Dynamic Layouts to effectively manage workloads

This also leads into Ab Initio’s Dynamic Layouts for HDFS, which gives you control, when workloads are heavy (or light) on the available computing capacity, for the degree of parallelism.

If you want to know more about Ab Initio on Hadoop, please check the great videos, white papers and documentation on the Ab Initio Discussion Forum.

## Side note

There is an underlying theme emerging of first, loosely defining the data, then subsequently applying processing steps to meet your desired business outcomes. This is evident in all the Ab Initio Frameworks we have discussed so far.

I believe the definition of the data should come from the Metadata>Hub. This would support good governance and ultimately an essential bias as to what data needs investigating in our digital enterprise. More on this later when we discuss the Governance, Audit and Control Layer.

## Want to know more?

If you want a full featured demo, then engage with your Ab Initio Account Manager. If you are in a knowledge repression culture, then drop me a line and I’ll help you connect.

In the next article, we wrestle the Real-time Layer within the [Ab Initio Lambda Architecture](https://www.linkedin.com/pulse/ab-initio-lambda-overview-chris-day-/).