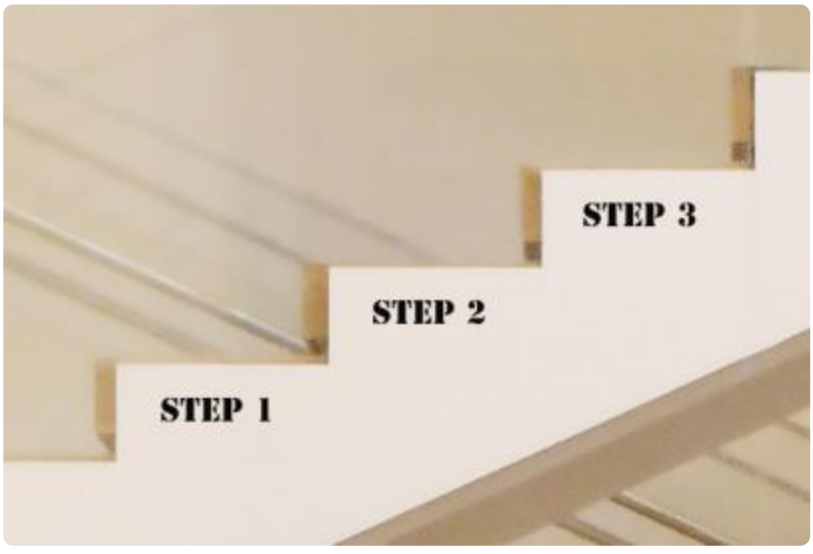
Building a Data Lake: Step by Step

Published on: October 29, 2019

Linda Feng, Software Architect



Top

At Unicon, a common challenge that we hear about is this: “We have lots of data everywhere, but we want a way to bring the data into one place so that we can analyze it and use it to ‘see’ how we are doing, possibly to spot trends and ultimately to inform decision making.” Many universities and school districts today are in various stages of implementing systems to enable data collection for useful analytics. A common problem is that they want to collect different types of data and combine them in meaningful ways. And while there are many more factors that contribute to the success of transformative uses of data on campuses, one hurdle that IT administrators face is how to get their “data house” in order.

During the last few years, I’ve spent most of my time helping customers assemble a variety of data sources into a data lake. What we have seen that works best is to first think through, as an organization, how you want to use your data. For example, what kind of information do you have in various systems, who are your stakeholders, and what information would they want to see? The next step is to organize your data into a central location. And whether or not you like the idea of a data lake or a data warehouse, or even a “lake house”, as one university administrator fondly joked, I’ve realized that having a system or checklist you can use to get your data house in order can be very helpful to guide the overall process and make things less overwhelming.

So why build a data lake in the first place? The benefit of a data lake approach is that raw data can be collected using whatever process is best suited for that source (i.e. REST API endpoint, CSV file uploads, database table copies, etc.). Key application data ingested into the lake can then be used to serve any number of downstream views, with optimized update capabilities based on delta changes. Perhaps most important, since the data lake will be used as the source for reporting and dashboard views, it will not impact normal operational database structures during normal business hours.

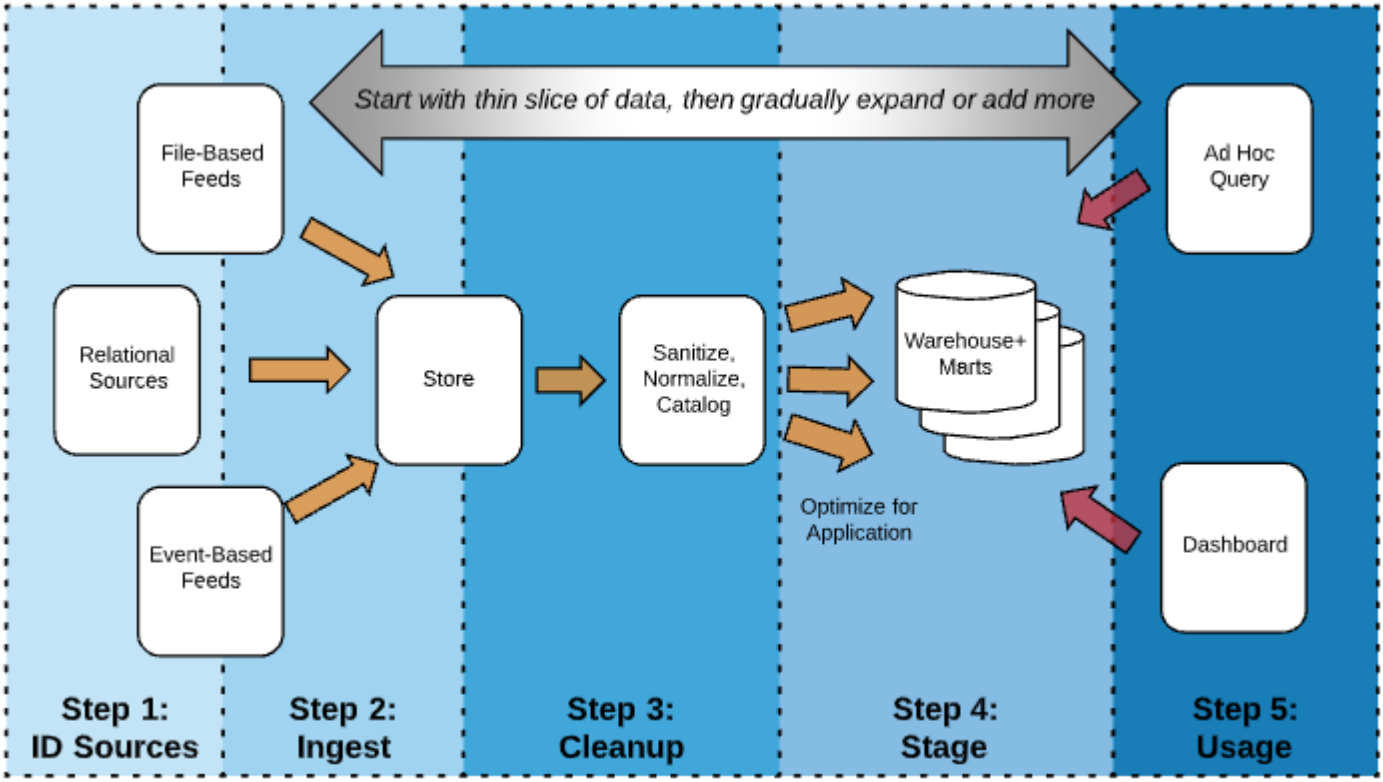
Not all components of this architecture need to be in place from the beginning; rather, they can be introduced iteratively over time. For example, during the initial phases of the project, data ingestion may be as simple as using simple storage, like AWS S3 and relational data stores, to store whatever available metric data might already exist, and exposing it “as is” to a front end dashboard API service. Down the road, when new requirements are introduced, this approach provides the flexibility to reconfigure data previously collected as often as necessary.

In my experience, having a checklist provides a methodical way to plan the necessary steps, obtain buy-in from the right stakeholders, and avoid common pitfalls.

Top

Checklist approach

This approach outlines a list of repeatable steps to identify key decisions and tasks that can be applied to each type of data as it is “hydrated” into the data lake.



Note that this checklist is laid out in a series of steps. However, it is possible and even recommended to carry out the steps iteratively, in a minimal fashion. For example, instead of getting all the data for a particular source, consider just getting a subset of the fields or rows. This will help you to go through the whole process and build up the infrastructure little by little along the way.

Some additional points to think about as you plan to go through the process:

- ✓ Avoid personally identifiable information (PII), if at all possible. This includes names, social security numbers, date of birth, passwords, etc. Instead, focus on obtaining the data via opaque identifiers.
- ✓ It is important to communicate to any outside departments/groups and stakeholders that will be impacted by any views of the data collected.
  - Ex: Executive Level Leaders, Students, Registrar’s Office

Step 1: Data Source Identification

Top

Data Source Identification is an important step that will be needed for every new type of information that needs to be collected. It is mostly an analysis task and usually involves a fair amount of inter-departmental communication.
Questions to ask:

- ✓ Is the data tracked in log files?
- ✓ Is it coming in a batch?
- ✓ Is the data generated in an event stream, meaning each activity is sent separately as it happens in the source application?
- ✓ Are there existing data stores that might be relational/structured?

For each identified source, you will need to set up access to the data origination environments.
Questions to ask:

- ✓ Who are the stewards/owners of the data origination environments?
  - Example: SIS (owned by the Registrar's or District Administration Office)
- ✓ Determine what data you actually need and communicate the specific needs to the data stewards/owners.
- ✓ Establish a process/plan for obtaining the data needed (both an immediate plan and a future plan).

Step 2: Data Ingest

Data Ingest is a largely technical effort but may also involve mapping to standard event record formats.

- ✓ For batch data, set up processes to schedule periodic file transfers or batch data extracts.
- ✓ For event data, set up processes to ingest the events - this might be an event endpoint. Ideally, if there is a standard event format (e.g., actor, action, object), you can set up a receiver function that will transform all the inbound events into the standard format prior to sending them through the data lake firehose. One example of a standard event format is IMS Caliper, which defines a framework for the collection and transport of learning events. Think of Caliper as a way to capture all of the interactions that a user has with a particular system or collection of systems.
  There are other event standards, such as Experience API. Caliper differentiates itself from similar standards by defining specific profiles for various types of learning interactions including reading, assessment, discussion and more. Additionally, Caliper has been developed with interoperability in mind and builds upon existing IMS standards including IMS LTI.
- ✓ For log data, determine how long it will be available. In some cases, it may be set to expire after a two to three week time period, and you will need to ensure that the full log history is

preserved.

- ✓ Set up the storage location, for example, an AWS account with S3 buckets, to serve as the data lake. To ease management, it is helpful to establish a consistent bucket naming and storage approach.
- ✓ Consider how you will deal with production/dev/test environments for your source and lake environments.
- ✓ Set up processes to bring in reference data (users, departments, calendar events, work project names).
- ✓ Consider other groups/departments that may be impacted by any new processes established and communicate the changes proactively.

Step 3: Data Cleanup/Organization

In the Data Cleanup/Organization step, you will deal with ways to combine data in more meaningful ways to serve downstream reporting/dashboard queries.

- ✓ Identify and locate common identifiers across the incoming data records
- ✓ Identify mappings between similar but differently named data fields and define logic for any transformations (parsing out specific identifiers from string fields, for example).
- ✓ Determine how to handle display fields that contain strings that might be too long or have unsupported characters
- ✓ Some data records may not contain an identifier that can be used to commonly join them across sources. In those cases, you will need to maintain a set of mapping tables with locally-sourced identifiers and their mappings to global identifiers.
    - This may mean that you need to manufacture a global set of identifiers to unify the data across systems.
- ✓ Coordinate and communicate with any departments/groups who own or can help locate the "source of truth" for identifiers.
    - Ex: UserIDs, StudentIDs, etc.

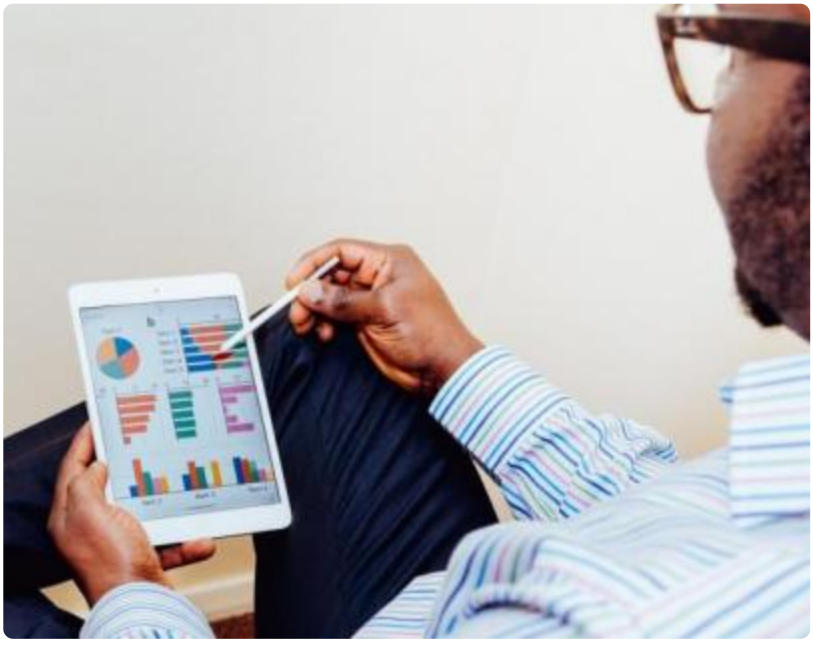
Step 4: Stage Data for Queries

This step enables data to be positioned into structures that are optimized for downstream usage. It is important to note that from the same data lake, different data "marts" can be positioned to serve a variety of downstream use cases.

- ✓ Consider the types of queries that will be needed for the data. This may involve working with departments to identify KPIs/important metrics that stakeholders need to know and that will help them make decisions.

Top

- ✓ Set up table layouts in the data lake.
- ✓ In some cases, it may be useful to aggregate metrics at logical boundaries.
    - Ex: usage stats for a day, week or month.
- ✓ For performance reasons, it may be useful to store the same data in different formats based on how it may need to be commonly displayed or accessed. For example, you might create multiple name fields: LastFirst, FirstLast, LastFirstInitial, and FirstOnly for different ways to search and display this type of data in front end apps.
- ✓ Consider taking advantage of serverless facilities that let you write SQL queries directly against files in cloud storage (i.e. AWS Athena or Google's BigQuery). If there is a small amount of frequently changing data that you want to use as a lookup dimension or if there is a large amount of data, it may make sense to do one pass through the data and store it in a cleaned-up format.
- ✓ Build out a library of queries that will be useful for dashboards and reports but that could also be used for ad hoc queries. For example:
    - Top 10 most active (users, systems, etc.)
    - Usage (last week, last month, etc.)
- ✓ Evaluate map/reduce tools: in some cases, it can help to avoid much of the details/complexity of big data processing - you develop your rules/code, let the map/reduce service manage jobs, memory, compute resources.
- ✓ Communicate these libraries to the necessary departments/users who will benefit from them. Documentation in the form of a data dictionary will be very important in order to help end-users understand how to interpret the data that they are querying.
    - Ex: Executive Level Leadership to gain insight into any additional query needs

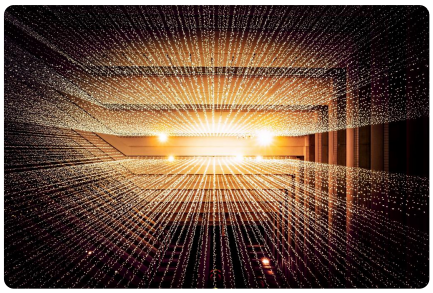


Top

Step 5: Visualize Data via Business Intelligence (BI) Tools

Once the data is staged, it can be accessed in various ways by multiple front end business intelligence (BI) tools.

- ✓ For commonly used BI tools such as Tableau, consider setting up workbooks with prepopulated queries or table definitions.
- ✓ Establish a BI environment for ad hoc queries.
- ✓ Evaluate whether queries and visualizations can be stored in Jupyter Notebooks, which enable sharing and reuse. Connect with data science professionals to prototype and validate algorithms.
- ✓ Work with key stakeholders to put together some preliminary dashboards and then run it through user testing to ensure that the views of the data are understandable and useful. It might be helpful here to research other dashboards in the industry to see what works and does not work for you.
- ✓ Maintain regular communication with the user community in order to determine new requirements for new or extended data sources.



A data lake is the best way to organize data from a variety of sources so that it can be analyzed and used to spot trends and inform decision making. Combining disparate sets of data is challenging, but taking the process one step at a time and having a set of tasks or checklist items for each step can make the process easier. It allows your team to be more focused and better able to coordinate with external groups, such as data owners for each of the source systems. When your data lake is built and your query and reporting tools are complete and functioning, it will be well worth it!

Additional Reading

Unicon Case Studies

North Carolina State Improves Student Success with Learning Analytics Technology (https://www.unicon.net/insights/case-studies/north-carolina-state-improves-student-success)
Jisc & Creating a Nationwide Learning Analytics Dashboard (https://www.unicon.net/insights/case-studies/Jisc%20and%20Creating%20a%20Nationwide%20Learning%20Analytics%20Dashboard)
Cloud Analytics Support National Course Outcomes (https://www.unicon.net/insights/case-studies/cloud-analytics-support-national-course-outcomes)

Top

Facilitating Individualized Assessments for Learning Analytics Readiness (https://www.unicon.net/insights/case-studies/Facilitating%20Individualized%20Assessments%20for%20Learning%20Analytics)

Data Lakes and Analytics on AWS

What is a data lake? (https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/)

Google Cloud Solutions

Cloud Storage as a data lake (https://cloud.google.com/solutions/build-a-data-lake-on-gcp)

Acknowledgements

Special thanks to Grant Huber for his assistance on Steps 4 and 5. Grant is a Digital Marketing Analyst for Unicon, and has completed projects implementing tracking and creating reports for Unicon's marketing website and other marketing funnels. Grant currently works as a consultant to help schools understand user behavior of on-boarding applications through the use of tracking software and dashboards.

Consulting, Digital Services, Integration, Cloud Services, Analytics, AWS, Governance

Linda Feng

Software Architect

Linda Feng is a software architect at Unicon, Inc., a leading provider of education technology consulting and digital services. Linda has deep experience in student information systems (SIS) integration, open standards, and big data/ learning analytics, most recently as Senior Product Manager for Canvas SIS Integrations and Canvas Data at Instructure. Prior to Instructure, Linda held the position of software architect for Oracle's Student Products Division. In the last several years, she served as co-chair of the IMS Global Learning Initiative Learning Information Services & Privacy Working Groups, helping to bring a new Enterprise interoperability standard to market.

Top

Insights and News

+ Articles (/insights/articles)
+ Blogs (/insights/blogs)
+ Case Studies (/insights/case-studies)
+ News (/company/news-media)
+ Events (/company/events)

Related Content

What Can an Instructional Designer Do for Me? (/insights/blogs/what-can-an-instructional-designer-do-for-me)

Unicon pledges to support data interoperability within K-12 education (/company/news-media/project-unicorn-vendor-pledge)

Agile in Real Life (/insights/blogs/agile-in-real-life)

Unicon Releases Federation Gateway for Streamlined Integrations to Okta (/company/news-media/unicon_releases_federation_gateway)

Unicon Pledges to Support National Cybersecurity Awareness Month 2019 as a Champion (/company/news-media/ncsam_2019)