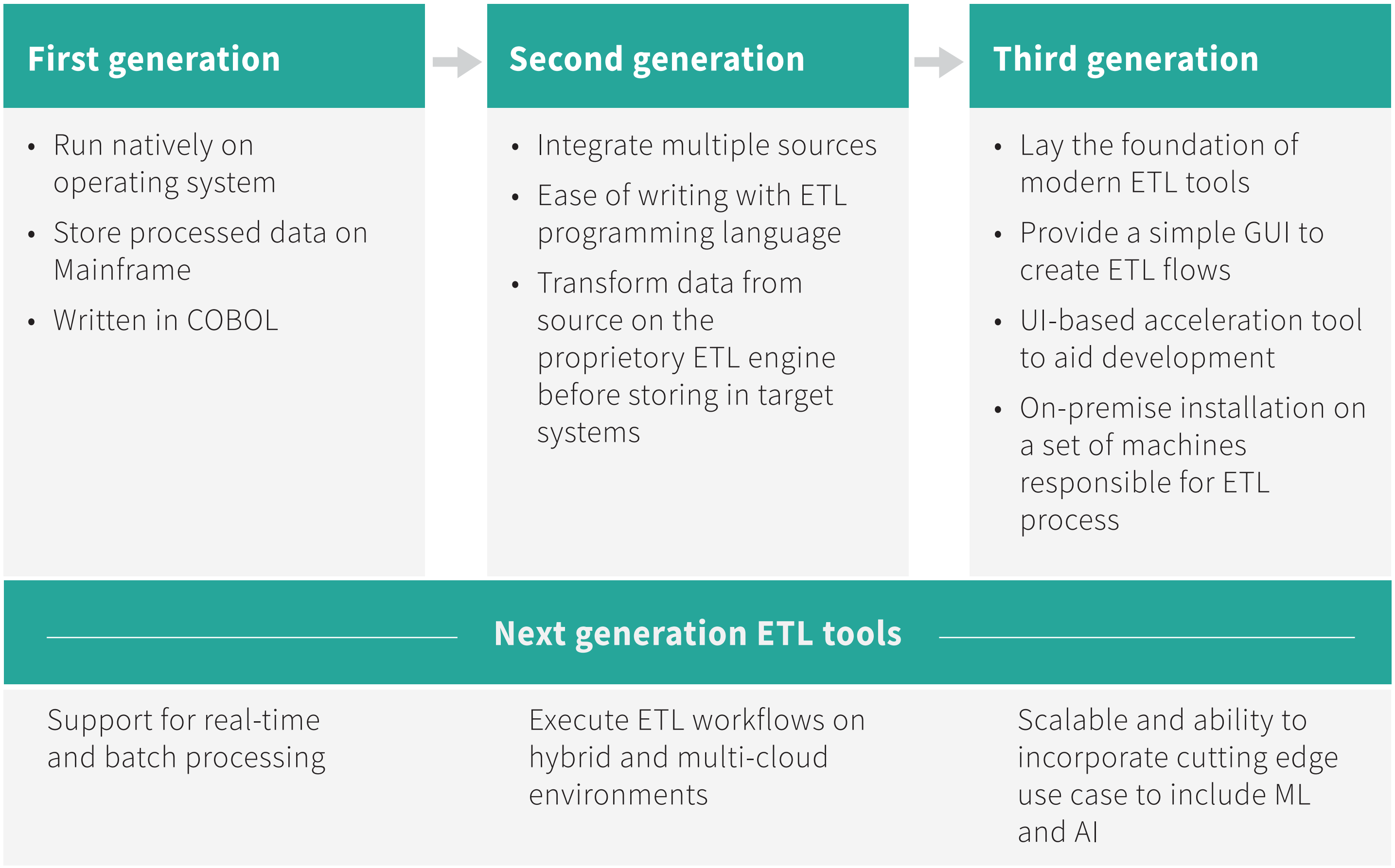
Modernizing legacy ETL platforms

Automate and accelerate your ETL migration to Spark in the cloud

Enterprises extract, transform, and load (ETL) data from multiple sources and applications to create a single data repository, a.k.a. data warehouse. ETL allows enterprises to effectively design and create an environment to mine and analyze data for making informed decisions. It isolates data from transactional systems, which ensures business-as-usual while data is analyzed in an optimized environment.

Evolution of traditional ETL tools

Traditionally, the ETL process involved building data pipelines in batches on-premise with limited sources, hardware infrastructure, and scope for real-time event processing. The tools have evolved over generations to cater to the changing needs of enterprises:



First generation

- Run natively on operating system
- Store processed data on Mainframe
- Written in COBOL

Second generation

- Integrate multiple sources
- Ease of writing with ETL programming language
- Transform data from source on the proprietory ETL engine before storing in target systems

Third generation

- Lay the foundation of modern ETL tools
- Provide a simple GUI to create ETL flows
- UI-based acceleration tool to aid development
- On-premise installation on a set of machines responsible for ETL process

Next generation ETL tools

Support for real-time and batch processing

Execute ETL workflows on hybrid and multi-cloud environments

Scalable and ability to incorporate cutting edge use case to include ML and AI

Note: ETL (Extract-Transform-Load) and ELT (Extract-Load-Transform) are often used interchangeably, based on the order in which the operations occur.

Challenges of traditional ETL tools

Time-consuming

Multiple components like metadata, custom sources, EDW, data marts, etc. need to interact with each other to create jobs

Expensive

High cost of ownership, operations, and maintenance

Lack of integration capabilities

Do not easily connect to existing infrastructure components, resulting in developing use cases from scratch

Limited transformation capabilities

Provide a limited set of transformations, making it difficult to customize flow design

Non-agile

Unable to build a use case with a limited subset of data and extrapolate it with the production dataset

Unscalable

Have fixed nodes and are not built to handle sudden spike in events

Handling failures

Use error-prone methodologies, which often result in data loss. Moreover, with interdependent processes, issues cascade across processes

Auditing

Cannot track changes made by users, making it difficult to find the root cause of errors

Licensing

Stringent licensing terms make it difficult to customize the platform for business requirements

To address these challenges, data-driven enterprises are shifting to next-generation ETL tools, which can run workloads on-premise and in the cloud. Unlike traditional ETL tools, these modern tools can extract value from extensive datasets. They also leverage the cloud without compromising security and provide better value for money.

Migrating to a modern ETL platform

A successful migration involves seamlessly porting existing ETL workflows to a new environment within the stipulated budget and time, without impacting business processes, for better performance. To ensure a successful migration, follow 3 I's – ideate, implement, and improve.

IDEATE

IMPLEMENT

IMPROVE

1. Ideate

Ideation is the most important step in planning a migration. Some factors to consider during ideation are:

- Available tools that can help with the migration
- Deciding whether to run the migrated jobs on-premise, in the cloud, or choose a hybrid strategy
- Identifying critical data that needs to be moved on priority vs. legacy data
- Impact on dependent applications
- Deciding on the sync interval
- Enforcing new compliance requirements like GDPR and CCPA

2. Implement

ETL migration involves moving multiple workloads without disrupting the existing workflows. Enterprises need to ensure the availability of data during migration for business-as-usual. After migration, data should be tested in the new environment to ensure it works with all the existing ETL workflows.

3. Improve

After the ETL jobs are migrated to the new environment, new workflows can be improvised, existing workflows can be enhanced by combining multiple redundant flows to save CPU cycles, and unnecessary jobs can be dropped.

Strategies to migrate to a modern ETL platform

Migration strategies differ across enterprises depending on their use case and operating environment. Businesses broadly have the following strategies to choose from:

1. Rebuild

All ETL workloads are built manually from scratch on the new system, like a hard reboot. The new tool also gives organizations the option to leverage additional features and overhaul the execution process. On the flip side, this approach is time-consuming.

2. Lift and shift

Lift and shift involve making an exact copy of the jobs from a legacy environment to the new environment. Each operator is stitched in the same order, and the logic is copied as-is. Once the right jobs are prioritized, identical workflows are created, allowing stable movement with minimal impact on existing processes.

3. Automate

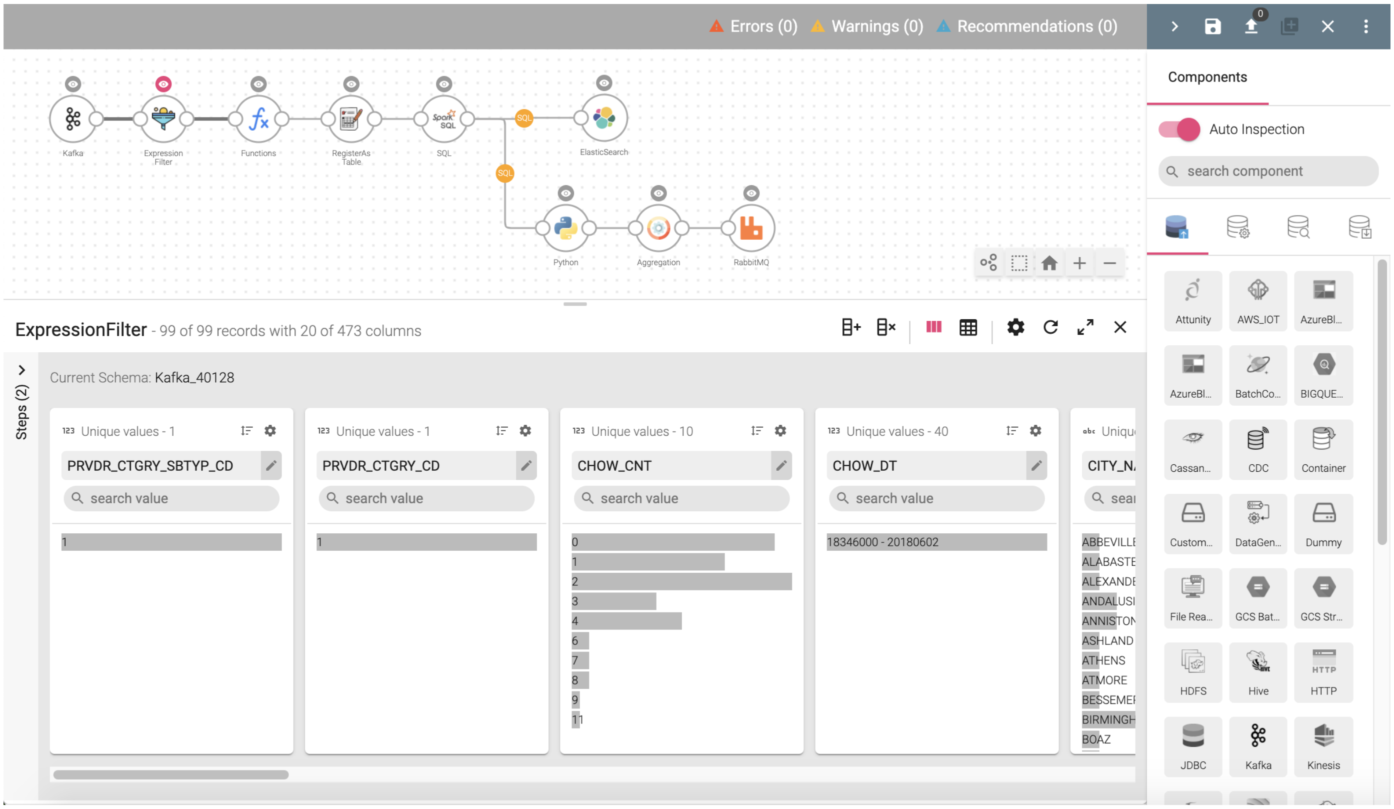
Automation is the fastest, risk-free way to convert existing workloads from traditional tools to new platforms. This strategy targets maximum automation and minimum manual effort. To ensure that the converted jobs work smoothly in the new environment, appropriate validation mechanisms are deployed.

4. Hybrid

Enterprises often adopt a hybrid approach – a combination of automation with rebuild/lift and shift – to ensure faster time-to-migration by cherry-picking techniques to rebuild or replicate certain flows in the target environment. This approach enables enterprises to fine-tune specific workflows, while most other flows are readily available in the new environment.

Modernize traditional ETL workloads with a self-service data flow and analytics platform

Gathr is a self-service ETL and analytics platform that lets you easily create batch and streaming ETL pipelines using drag-and-drop operators on a visual IDE. Gathr has a wide array of built-in operators for data sources, transformations, machine learning, and data sinks.



ExpressionFilter - 99 of 99 records with 20 of 473 columns

Current Schema: Kafka_40128

Gathr enables enterprises to build production-grade continuous applications with machine learning capabilities. It helps users derive maximum value from their data, maintain greater consistency within data streams, and join streams with static data sources efficiently.

Migrating ETL workloads to Gathr

Gathr provides a comprehensive environment for your migration needs from traditional platforms. It takes a three-step approach – assessment, conversion, and validation – to ensure flawless execution of the migration processes with limited room for disruption.

1 ASSESSMENT

Existing workloads are assessed for complexity, size, and compatibility operators

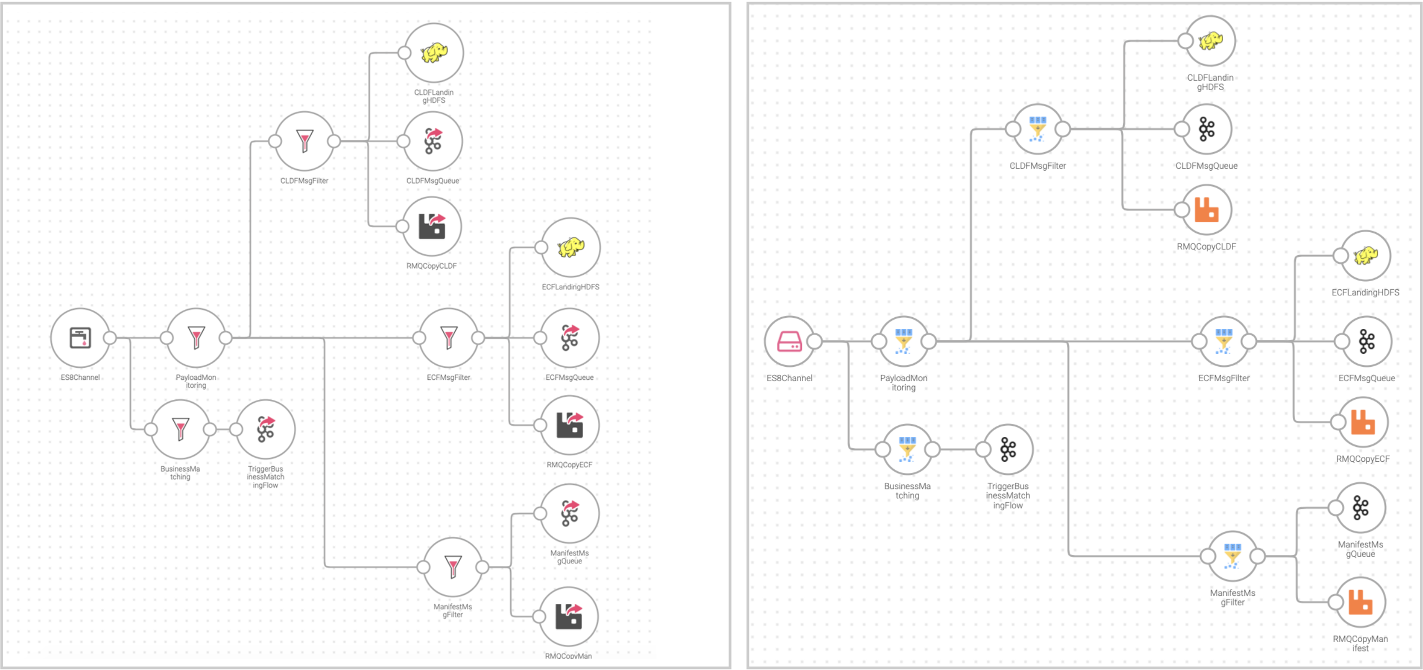
2 CONVERSION

Migration is executed using various migration strategies. Traditional ETL workloads are transformed into Spark-based distributed workflows

3 VALIDATION

Ensuring successful migration of existing workloads without data loss

This three-step process results in a fully functional combination of workflows and pipelines in Gathr, which is identical to the existing workloads. It also provides an integrated validation process to ensure a successful migration.



ETL workflow on traditional platform

An identical workflow on a modern platform

Migration from a traditional ETL tool to Gathr

Advantages of migrating to Gathr

1. **High performance**

   The migrated workloads run on a distributed high-performant architecture. With automatic support for Spark and native cloud execution engines, Gathr can boost the way ETL jobs are executed and process data in minimal time with limited resources.

2. **Elasticity**

   Gathr workloads can be configured to scale up and down automatically depending on the rate at which data is generated at the source. When creating jobs on the cloud, this helps control costs by provisioning resources only when required.

3. **Process massive datasets**

   To leverage cloud data warehouses, enterprises must focus on processing and ingesting data from multiple relevant sources. Gathr can process massive datasets from varied sources and push them to all major data stores for powering the analytical engine.

4. **One-stop solution**

   Gathr is a unified solution with powerful capabilities for building any type of ETL flow. Instead of procuring multiple products for data cleansing, transformations, profiling, monitoring, cataloging, and preparation, Gathr can perform these tasks end-to-end.

5. **Designed for cloud**

   Gathr provides extensive support for cloud-native services, including real-time and batch sources, and multiple specialized services to power data stores in the cloud. It can also run workloads on the native engine itself to maximize efficiency and cut down execution time.

6. **Complex orchestration**

   Many business use cases require creating multiple pipelines, which involves coordination between pipelines. It also introduces dependencies based on data or time, which requires the ability to orchestrate the inter-dependent flows. Gathr has an in-built capability that takes care of complex orchestration.

Beyond migration with Gathr

Once you migrate your existing workflows, you can further enrich them with next-generation capabilities of Gathr to onboard use cases.

Hybrid execution

01

Freedom to choose and execute ETL workflows on multi and hybrid cloud environment

Machine learning & advanced analytics

02

Ability to incorporate advanced analytics and machine learning capabilities to your ETL workflows

Integrate new sources

03

Integrate and process data not only from traditional sources like databases, but also from real-time and cloud-native sources

Specialized data lakes

04

Leverage advanced data lakes to power real-time insights

Easily build fast and reliable data pipelines using Gathr

Start for free

Gathr is a next-gen, cloud-native, fully-managed, no-code data pipeline platform. It's the only all-in-one platform for all your data integration and engineering needs – batch and streaming ingestion, CDC, ETL, ELT, data preparation, machine learning, and analytics. The Spark-based platform brings unmatched speed, performance and flexibility required to handle all types of data and analytics approaches, in ways that traditional ETL tools cannot. With Gathr's visual drag-and-drop interface, native integration for all popular data sources and destinations, an exhaustive set of pre-built operators, and a rich pipeline template gallery, anyone can build and deploy data pipelines, quickly and easily.

Visit www.gathr.one or write to us at contact@gathr.one