gathr

Power massive scale, real-time data processing by modernizing legacy ETL frameworks

Enterprises rely on actionable insights for making strategic business decisions, for which they need to analyze huge volumes of data from varied sources in real-time. ETL data flows can help address this need by consuming raw data from different sources, transforming it, and storing it for further analytics.

To power real-time decision making on large data sets, enterprises need an expert team, high-performing hardware systems, and a scalable ETL solution that can accelerate development and deployment of ETL frameworks, while swiftly accommodating changing business needs.

Enterprises that have experience in processing large data sets usually create custom frameworks to achieve near-term, non-functional goals and business requirements.

However, with changing business directions, technical debt, and dependency on IT teams who understand the historical choices made during the initial platform designs – organizations run the risk of impacting businesses and increased customization cost.

Next-generation ETL tools allow enterprises to effectively design and create an environment to mine and analyze data for making informed decisions. They isolate data from transactional systems, which ensures business-as-usual while data is analyzed in an optimized environment. These frameworks also help users solve business problems without spending cycles perfecting boilerplate code.

About the customer

A leading security and intelligence software provider focused on creating powerful intelligence and investigation technologies for federal and state-level security agencies. Their solutions enable the security agencies to understand the cyber threats through intercepting communication data, data integration, and advanced data analytics by leveraging artificial intelligence models on big data.

Business needs

The communication analytics solutions provider wanted to modernize their existing big data applications and was looking for an easy-to-use and scalable solution that could process 1.5 billion transactions (user interactions) generated per day from multiple real-time feeds.

They were looking for a near-zero-code solution for ETL processing jobs that would:

- Perform real-time ingestion and complex processing (parsing, enrichment, and aggregation)
- Ensure high throughput while indexing and storing
- Ensure high performance with low-cost commodity hardware
- Ensure conditional monitoring of data in real-time for reporting and validations
- Ensure high throughput for data-intensive aggregations
- Provide support for near-real-time and batch processing
- Support continuous integration and continuous delivery
- Detect anomalies in transactions
- Create additional data processing flows for compliance and regulatory reasons

Solution

Gathr enabled the client to implement applications that run on a scalable Spark compute engine as structured streaming data pipelines while providing self-service and analytics capabilities for large-scale data processing.

The ETL solution used Gathr's vast library of components for data acquisition, processing, enrichment, and storage. The entire data flow was created and orchestrated in Gathr's Web Studio using a low-code methodology.

The key technologies and components involved were as follows:

- **Kafka** to stream data in real-time
- **Gathr's out-of-the-box ETL components** for data processing:
  - Real-time data quality detection and alerts
  - In-memory data transformations like filter, evaluating static and dynamic fields, data type conversions, and message parsing
  - Enrichment of real-time data by looking up syndicated data marts and in-house technology stores such as Redis, HBase, and FTP server
  - Aggregations like min, max, count, and sum over a large moving time window of 6 hours
  - Monitoring and capturing data statistics on moving data
- Used **Gathr processors** and storage components to create Polyglot architecture:
  - Automatic tie-up field insertion
  - HBase to store the transactions
  - Solr to index data for BI queries
- **ClearInsight, a rapid application development platform** that enables agencies to create applications precisely designed to their investigative needs using low code/no-code methods

Lengthy development cycles

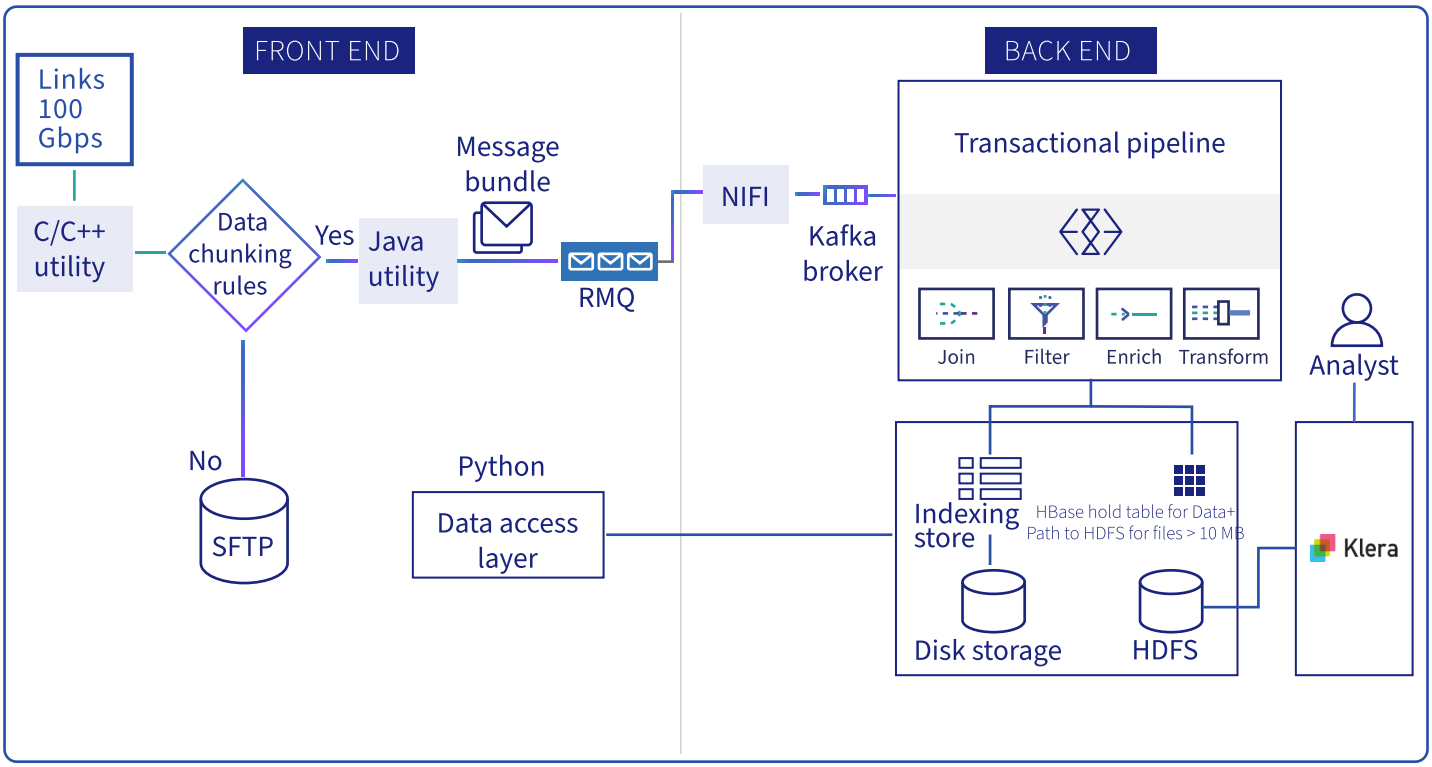
Inability to alter processing behavior via configuration changes

Time-consuming debugging and rectification process

Lack of version management and hot-swap features

Complex operations

Stringent SLAs on data availability and query results



ETL processing using Gathr pipelines

Solution highlights

- Drag-and-drop based visual user interface
- Data-driven design and testing features
- Single dashboard for debugging, managing, and monitoring the solution
- Continuous integration and delivery from development to QA to staging and production
- Design and schedule batch flows for periodic aggregations
- High performance with Spark structured streaming

Business benefits

Gathr enabled end-to-end data ingestion, enrichment, machine learning, action triggers, and visualization to modernize hand-written big data applications to Spark structured streaming in weeks. This, in turn, helped the communication analytics solutions provider realize several strategic benefits:

1. Replaced roughly ~1 million lines of code in ~3 weeks using Gathr frameworks
2. Achieved a high throughput of 100000+ transactions/second, enabling
3. Reduced the overall release cycle from 8 months to 8 weeks
4. Reduced the release cycle for new changes from 3 weeks to 3 days
5. Saved overall project cost by designing the solution on commodity hardware