Whitepaper

Accelerate Apache Spark Development and Operations

Build and run applications rapidly through a low-code approach

Introduction

Apache Spark, a fast in-memory data processing engine and a distributed computing framework, can handle all needs for batch and streaming data processing, analytics and machine learning. Large data-driven enterprises are using Spark for tasks ranging from ingestion, ETL, and data quality processing to advanced analytics and machine learning.

Despite its growing popularity, Spark is complex and the learning curve is steep. Developing code, integrating and testing with Spark is often time-consuming.

Gathr.ai offers a low-code solution to the complexities of building enterprise-grade Spark applications. It enables visual, built-in, workflows instead of manual programming. This reduces the time to develop and operationalize applications. The Gathr platform helps visualize data sources, data preparation, business logic, and third-party interfaces. It is a scalable enterprise grade platform and is accessible to both developers and business users.

Build and run Apache Spark applications rapidly using Gathr.ai

Intuitive UI

Simplify and accelerate Apache Spark development with an intuitive interface and a visual pipeline designer

Click + Code

Use 150+ drag-and-drop Apache Spark connectors and operators. OR, use native Apache Spark based operators and languages including Java, Scala, SQL, and Python for hand-coded custom logic

Full application lifecycle support

Supports the entire application delivery lifecycle: design, build, test, debug, deploy, monitor and manage

Why Gathr?

Gathr enables rapid Spark application development and end-to-end data processing. It facilitates a seamless way to operationalize and monitor application, and offers a time-to-market advantage of visual development vs. hand-coding

The platform includes one-click deployment options, application governance tools such as data inspection (iterative debugging), data lineage (run-time audit of the complete journey of a data record), and active alerting and monitoring capabilities thus enabling higher productivity across a complex set of tasks.

Technology Value

Enables accelerated Spark development	Pre-integrated drag-and-drop operators in a visual UI allow 10x faster development compared to manual coding
The only multi-engine real-time and batch analytics platform	The only platform that offers a visual UI-driven DevOps approach for Apache Spark, as well as other engines including Apache Storm, Apache Flink, TensorFlow and Oozie in a single platform
Data 360 capability	An end-to-end big data processing platform for data ingestion, ETL, analytics, machine learning, action triggers, and data visualization
Enables machine learning at scale on real-time data	Use a comprehensive set of advanced analytics and machine learning operators, such as Spark MLib, Spark ML, PMML, H2O, and Tensorflow. Train and score models across real-time and batch workflows
Offers ease of DevOps and monitoring	Run streaming and batch pipelines constantly and consistently once they are in production
Quick deployment to test/ production environment	Use one-click deployment options to deploy applications on-premise or on a public cloud
Version management	Create and save new versions of your data pipelines, and conveniently roll-back changes by reverting to an older version
Intelligent real-time alerting	Set thresholds on key metrics and corresponding action triggers through real-time alerts, enabling appropriate actions before performance and availability issues occur
Safeguards against technology flux	Work with the power and flexibility of best-of-breed open source technologies integrated into a high-performance, scalable, reliable and constantly evolving enterprise-grade platform

Enhanced productivity for Spark development and management tasks

The following table provides a snapshot of tasks where Fortune 500 companies are achieving significantly higher productivity with Gathr vs. manual Spark development.

	Manual effort	Gathr.ai	Productivity gain
Coding/testing of a simple 3-step pipeline (read, transform data, write)	3-4 hours	10-30 minutes	Minimum productivity gain: 6x Maximum productively gain: 24x An average of 10x speed vs. hand-coding across use cases
Coding/testing complex pipeline with data source blending, complex-event processing and time-window aggregations	3-5 business days	3-4 hours	Approximately 8x to 10x speed
Complex applications in multi-tenant mode	8-10 member team	2-3 member team	3x to 5x higher productivity without increasing team size Tasks are 5x to 10x faster
Real-time dashboards (build, configure, and deploy)	4-5 business days	30-60 minutes	30-40x speed
Minimal overall staffing for Spark DevOps team at the scale of a large enterprise for a shared services IT team	6-10 engineers	2-3 engineers	Save $700K to $1.4M per year for every 10 manual Spark developers

A deep dive into productivity metrics

Coding/development effort

Here are examples of productivity benefits that Gathr provides ranging from simple to complex pipeline development:

1. Coding and testing of a simple three-step pipeline read from Kafka, transform, write to HDFS

Gathr.ai	Manual
Method of development: Self-service, UI based, drag/drop pre-built operators. No deep Spark skills required	**Method of development:** Spark programming skills required, including Spark Streaming, Kafka, Hadoop and all inter-system interfaces
Steps: Drag/drop 3 operators, connect, configure, deploy	**Steps:** Coding each step; manual configuration parameters, build, compile, debug, test locally, and finally deploy on a cluster with appropriate shell access permissions needed
	Approx. Time: 3-4 hours

2. Coding and testing of a complex pipeline (read from Kafka), filter, enrichment and blending static and streaming data, complex event processing (CEP), and analytics

Gathr.ai	Manual
Steps: Drag/drop operator for each operation, configure, customize, visual debug with sample data, deploy	**Steps:** More varied and deeper skills required
Approx. Time: 3-4 hours	**Approx. Time:** 2-3 days

3. A real-world scenario – coding and testing of multiple complex applications on a common data platform in a multi-tenant mode

Gathr.ai	Manual
2-3 member team to develop and operate. Each use case will take a few days to deploy	8-10 member team - with most tasks taking 5-10x longer to achieve

Data visualization

Real-time dashboards (build, configure, and deploy)

Built-in and custom real-time dashboards display the status of metrics and key performance indicators for a pipeline. Gathr.ai allows you to blend real-time and historical data with offline analytics and integrate everything you must keep track of, however disparate, onto a single screen.

Gathr.ai	Manual
Method of development: Built-in feature. Anyone can create and access a real-time dashboardno coding skills required	**Method of development:** Needs to be built from ground up. Needs expertise in HTML5, web sockets, graphs/charts and other technologies
Steps: Configure and build Approx.	**Steps:** Coding each step; manual configuration parameters, build, compile, debug, test locally, and finally deploy on a cluster with appropriate shell access permissions needed
Time: 30 minutes	**Approx. Time:** Weeks of development for a full-fledged dashboard for each application

DevOps and monitoring

Gathr simplifies DevOps, monitoring, and error handling tasks with several added benefits over manual coding across operations such as: movement of applications, version management, application performance monitoring, alerting, data inspection and lineage.

Manual coding requires building each capability and feature from the ground up, which can take a few weeks to several months. It further needs manual binary transfer, deployment, reconfiguration, testing/ certifying, which can be highly error prone.

DevOps and monitoring tasks	Gathr.ai advantage
Promotion and movement of applications among different environments	UI-based easy movement of code/ applications from development to test to production environments
Version management of applications	• Built-in feature with visuals of multiple versions: Create and save new versions of your data pipelines • Roll-back changes conveniently by reverting to an older version with one click

DevOps and monitoring tasks	Gathr.ai advantage
Application performance monitoring and alerting	Built-in feature with visuals of multiple interactive graphs / statistics / metrics and alerts
Data lineage and inspect	Use the built-in data inspection feature during development, and data lineage for your production deployed pipelines: • End-to-end view of data transformation before and after the use of every operator • Capture and record data changes at every stage in the pipeline • Search and view the entire path of incoming messages through the pipeline, including the processing time at each stage and changes to the data attributes

Moving applications to production

Use one-click deployment options to deploy applications on-premise or on a public cloud. And, constantly run streaming and batch pipelines once they are in production.

Built-in features:

- Infrastructure failure response: Ability to retry multiple times in case of infrastructure issues (such as intermittent network connectivity or service downtime)
- Error management: Capture logic errors (such as parsing issues, null-Pointers, etc.) and enable analysis
- Monitoring and alerting: Monitor and alert on processing throughput, processing time, slowness (queued batches), and pipeline crashes
- High availability: Gracefully recover from a node down scenario
- Log aggregation: View and search logs for useful information across all nodes on the web UI
- Data lineage and audit: Track data changes during every processing stage in the pipeline(s)

Examples of customer use cases

Below are examples of the effort involved in building pipelines ranging from low, medium to a high level of complexity, using Gathr. These efforts include understanding requirements, installation, development, testing, performance benchmarking, and the time to deploy the pipelines into production.

Use Case	Complexity	Data Rate	Effort*
1. 7 pipelines built	Medium to High: Ingestion from Kafka, transformation, aggregation, co-relation, persistence (Hive, Elasticsearch), reporting (Kibana, Tableau using Presto)	130 million events/day ~370GB/day	8 weeks (for all 7 pipelines) – 2 people
2. 12 pipelines built	High: Ingestion from Kafka, transformation, CEP, analytics, persistence (HDFS, Hive, Elasticsearch), reporting (Kyvos, Kibana)	80 million events/day ~4 TB/day (designed to scale up to ~40TB/day)	10 weeks (for all 12 pipelines) – 2 people
3. 3 pipelines built for Call Centre Analytics	High: Ingestion (from Kafka, Syslog, log files), filtering, out of order events (late arrival), CEP, alerting, aggregation (over 30 minutes using external cache), persistence (Elasticsearch), reporting (custom web UI)	50 million events/day ~300GB/ day	6 weeks – 1 person
4. 2 pipelines built for insider threat detection	High: Ingestion (from files and Kafka), filtering, de-duplication (over 7 days data using external cache), insider threat detection model, alerting, persistence (SQL Server and HDFS)	~400 million events/ day (scale testing done for 1 billion events/day)	6 weeks – 1 person
5. 3 pipelines for CDC use case from Oracle	High: Ingestion (from Kafka using Attunity and HDFS), transformation, Hive(ORC format), workflow (Oozie based)	~100 million events/ day	4 weeks – 1 person

About Gathr.ai

Gathr.ai is an enterprise-grade visual platform for all your streaming, batch data processing, and analytics needs.

It allows you to ingest, blend, and process high-velocity big data streams as they arrive, run machine learning models, visualize results on real-time dashboards, and train and refresh models in real-time or in batch mode.

You can now build and operationalize custom big data applications five to ten times faster using a visual drag-and-drop interface, an exhaustive set of pre-built operators, full application lifecycle support, and one-click options for on-premise and cloud deployments.

With support for multiple big data engines and functional extensibility built-in to the design, Gathr gives you full flexibility and control to work with the technology stack of your choice.

Scan and
start free 14-day trial

Scan to
schedule a demo