Data Factory
DRAGONFLY
Mine, Manage and Monetize Big Data

Dragonfly DataSwarm™ & DataSwarm Marketplace Differentiated Value Proposition

Table of Contents

DataSwarm Solution Differentiation

For most businesses, the Internet of Things (IoT) brings three fundamental challenges: (1) handling the tsunami of data coming from sensors and smart devices, (2) detecting and responding to significant events as fast as possible, and (3) providing an integrated view of historical and current business performance. Meeting those challenges requires a scale-out Big Data approach to real-time event processing, and mechanisms for integrating data and analysis across batch and real-time domains.

Technologies such as Apache Storm provide a robust programming environment upon which data scientists and engineers can build real-time applications. However, those applications must be programmed, deployed, managed and integrated by hand. The amount of manual crafting required can come as a shock to organizations accustomed to the levels of abstraction and automation provided for years by traditional data analysis tools.

That's where Dragonfly Data Factory™ comes in. We've developed Dragonfly DataSwarm, a real-time IoT analytics and automation platform designed to:

Accelerate real-time application development

- **Rapid GUI-based composition of applications –** from data-source, analytics, sink and action components. Out-of-the-box components include real-time sources/protocols (Kafka, MQTT, AMQP, XMPP, Kinesis); ETL functions; analytics; and alerts.
- **Component framework (SDK) –** enables easy creation, management and reuse of additional components.
- **Multiple stream analytics execution-logic formats – i**ncluding SQL, Java, and Predictive Model Markup Language (PMML).
- **Global online component library –** constantly updated by Dragonfly and an ecosystem of developers.
- **Local online component library –** for sharing and reuse of components within an organization.
- **Built-in analytics dashboard –** with an SDK for creating new display components.

Integrate real-time and batch-based analytics

- **Data integration –** enable real-time use of batch data, and batch use of device-generated event data, including real-time query.
- **Processing and data orchestration –** update static repositories and/or trigger batch-job execution in response to real-time events (received or derived by analytics); dynamically update in-memory reference data used by real-time applications based on source-data change events.
- **Integrated visualization –** customizable displays, mixing real-time events and responses with historical data views, trends, inflection points, etc.

Simplify real-time application deployment lifecycles

- **Application management across execution-engine clusters –** Deployment, un-deployment, rebalancing, starting, stopping, monitoring, etc.

- **Execution-engine cluster resource management –** Add/remove nodes, performance monitoring, etc.
- **Model-based, preemptive tuning and outage avoidance –** monitoring and predictive analysis of real-time application and infrastructure performance.
- **Non-disruptive updates –** Adjust running-application component parameters, update running-application component code, and upgrade the DataSwarm platform without any application down-time.
- **Multi-tenant support –** Secure separation of platform users and their data.

Dragonfly DataSwarm enables organizations to shift their focus from software development to data science and analysis, accelerating the delivery of analytics project value—and marketplace responsiveness.

DataSwarm Overview

The emergence of the Internet of Things (IoT) and the constant exponential growth of data emission by sensors, machines, vehicles, mobile phones, social media networks, and other real-time sources are compelling organizations to rethink their data and analytics strategy beyond batch-processing. They are increasingly aware of the need to have access to the latest information to gain a competitive advantage.

Open source community is offering cutting edge innovative technology to address this need, but however, building solutions using open source from scratch may be expensive and time-consuming. DataSwarm eliminates this problem and provides a platform powered by open source engines with useful features, flexibility, extensibility, ease-of-use, and monitoring support.

Real-time decision support and analytics use-cases, today, are best optimized by utilizing different stream processing with low latency and event level processing, whereas in other cases, the micro-batch computation is the best fit. DataSwarm simplifies the trade-off by integrating multiple engines in a single platform, and eliminating the whole effort of integrating the different underlying technologies.

DataSwarm is the multi-engine platform with support for Apache Storm and Apache Spark Streaming that offers a flexibility to execute data pipelines using a stream processing engine of choice - to eliminate the pain of dealing with multiple frameworks separately, each built for a niche purpose.

DataSwarm is designed to continuously ingest massive volumes of data, to rapidly build and deploy streaming analytics applications for any industry vertical, any data format, and any use case. The high performance stream processing engine continuously queries, filters, correlates, integrates, enriches, and analyzes data to discover exceptions, patterns, and trends that are presented through live dashboards.

Elements of DataSwarm

DataSwarm is a powerful and scalable real-time analytics solution designed to provide rapid application development through a rich graphical user interface. DataSwarm abstracts the complexity of programming real-time applications with its high degree of automation using a powerful drag and drop GUI. It enables the organizations to shift their focus from software

development to data science and analysis, accelerating the time to market in the analytics value chain.

In an application topology, built on DataSwarm platform, **Components** are the fundamental building blocks of the platform.

Components could be of the following type:

- Sources
- Processors &
- Sinks

Sources read data from raw data sources like Streaming API's, Apache Kafka queue, Kestrel queue, MQTT and so on.

Processors are the logical processing units. Processors receives input stream from sources, process and emits the output stream to another Processors or sinks. Processors can perform the operations of filtering, aggregation, joining, interacting with data sources and databases.

Sinks are the output data sources where the processed data will be written to. A processor processes the stream and writes the output data to any of the following sinks

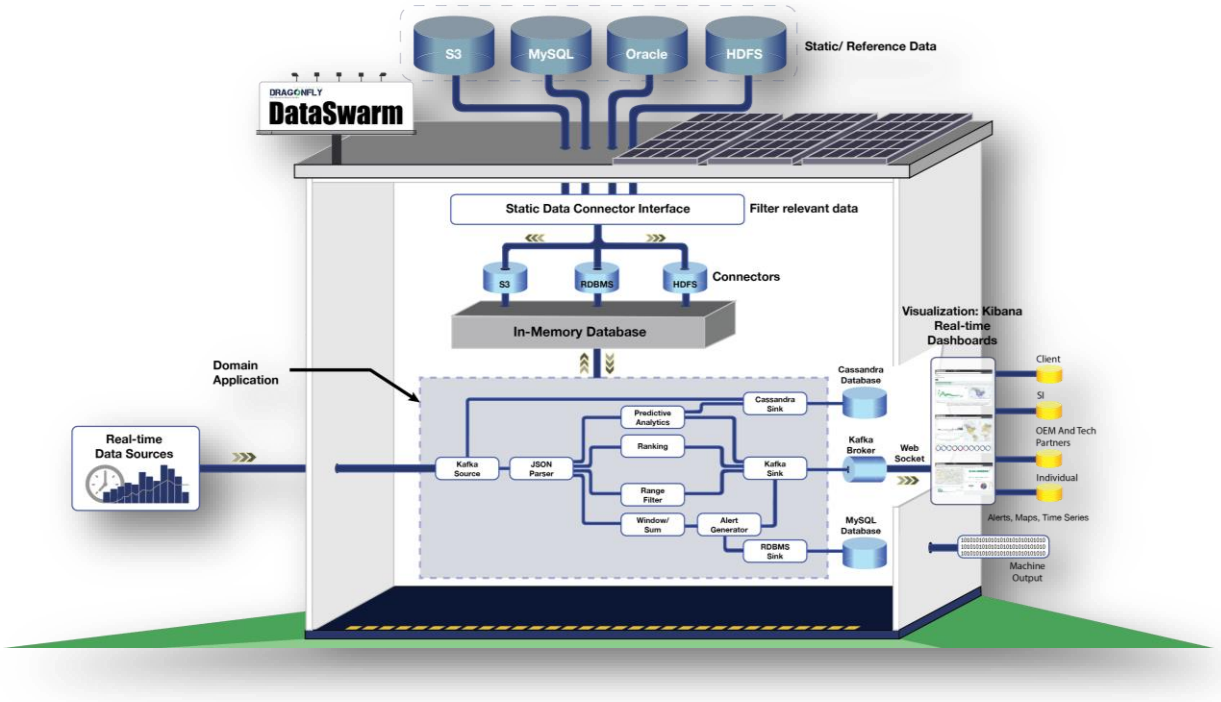
- Message Queues like Kafka, MQTT
- NoSQL databases like MongoDB, Cassandra and so on

Sources, Processors and Sinks are connected together to form an Application or Application Topology. Real-time application logic is specified inside the DataSwarm topology. In simple words, a topology is a directed graph where vertices are computations and edges are stream of data.

A simple Application starts with Sources which emits the data to one or more Processors. Processors represent a node in the topology having the smallest processing logic and the output of a Processor can be emitted into another Processor as input or can be written to any of the sinks.

For referencing and contextualizing the data, Users can load the static/reference data, from any data source to the In-memory databases like Hazelcast using **Connectors**.

DataSwarm Solution Architecture



DataSwarm processes the real-time transactions data, and loads the static/reference data onto the In-Memory Database (IMDB), from various relational databases like – S3, MySQL etc., and can be scheduled using connectors.

Using built-in processor component, DataSwarm can do the contextualization, and integrate different static data in IMDB with the real-time stream, using primary key concept.

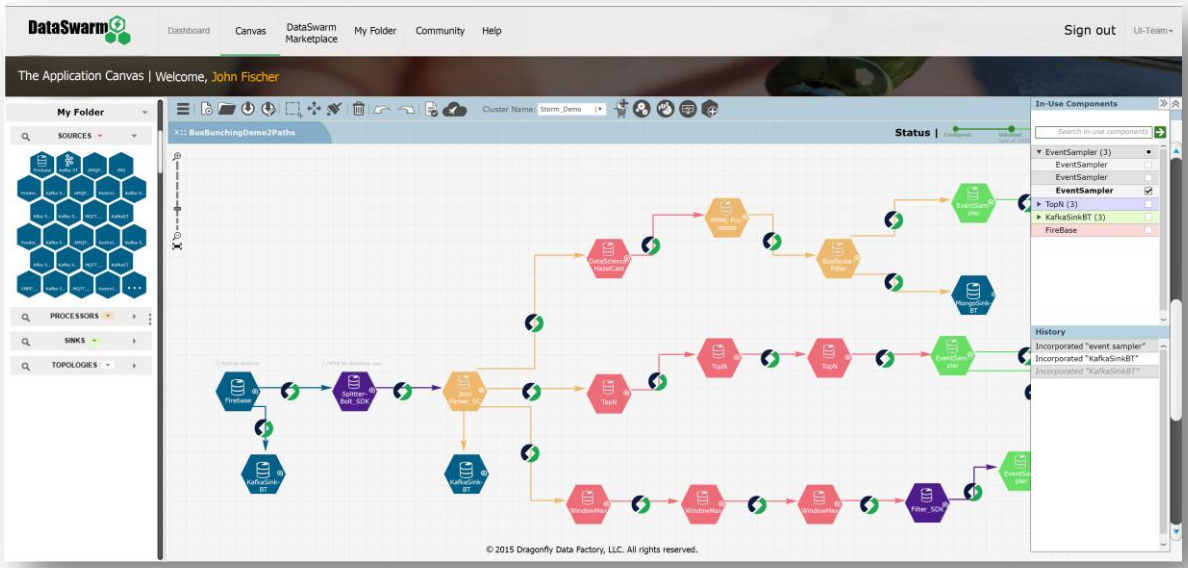
Processing, analysis and predictions can be performed on the data, using built-in components and machine learning models. The output and results of the application, can be stored in different databases, for different purposes, either for visualization (static/real-time), or for triggering and sending alerts to machines/sensors.

DataSwarm Features

Simplified Real-time Application Building Process

- Building stream processing applications can be time consuming and complex. DataSwarm dramatically simplifies this process with a friendly UI and a rich set of pre-built components. Developers can use the visual interface to configure and deploy new applications in minutes with minimal or nil custom coding required.
- DataSwarm provides a rapid graphical composition of applications from source, analytics, sink, and action components, and users can develop and deploy applications on Apache Storm and Apache Spark streaming based on the choice, without reprogramming the whole architecture.

- Using software development kits, developers can easily create, manage and reuse the components for application composition



Software Development Kit (SDK)

- Dragonfly provides a simple component development framework i.e. SDK, for developers, for rapid component development. It comes with multiple stream analytics execution logic formats like - SQL, Java, and rule-based (most likely JBoss) to initiate complex event processing.
- Using SDKs developers can simply create organization specific business logic (e.g. custom components) from any data source and make it available for data scientists to visually assemble the streaming application.
- SDK also provides user created GUI-configurable custom component support

Visual Designer Interface

- Create real-time streaming data analytics applications in minutes with a powerful visual editor
- Design, import and export real-time data pipeline
- Drag, drop and connect components to create applications
- Easily connect different pipelines together with sub-system integration
- Monitor detailed metrics of each task and each instance
- Provides seamless integration with Indexing store (e.g. Solr, Elasticsearch) and NoSQL databases (e.g. MongoDB, HBase, Cassandra etc.) for writing data in real-time

Real-time Platform

- Facilitate rapid development of industry specific solutions with complex real-time applications and pre-built components

- Define message schemas and alert rules for a real-time application
- View, edit, modify, export, import the entire real-time pipeline
- Provides operational insights of a running application, like information on its throughput and mean processing time
- Developers do not have to worry about multi-threading the code, the application is automatically partitioned and distributed across the Hadoop cluster for scalability.

Cluster Monitoring Tool

- Visually allows operations personnel to provision, manage, and monitor the entire cluster
- Installs and deploys all the underlying technologies/ libraries required for running DataSwarm in a cluster mode with a few clicks
- Monitor detailed metrics of each task and each instance

Real-time Analytics

- DataSwarm increases the efficiency of the application by leveraging drag-and-drop components for predictive and analytical modeling for live predictions
- Can run PMML-based scripts in real-time on every incoming message
- Connectors will help to blend streaming data with static data without any coding, for contextualization and resilient predictive modeling

Real-Time Dashboard

- The dashboard will show a visualization being automatically updated with the latest data. Basic descriptive graphs can be built and included in the dashboard, to monitor multiple metrics in a unified view
- The real-time dashboard enables the organization to take immediate data driven decisions, based on the trend and the customer behavior, and also enables them to monitor the Key Performance Indicators (KPI) of the business on a real-time basis.
- Processed output data is passed to ELK stack, for generating real-time dashboard using open source visualization tool - KIBANA.

- Platform can be easily integrated to any real-time or batch visualization tools.

Data Integration

- DataSwarm will enable real-time process to write data to RDBMS or Hadoop repository
- Static data can be blend with the real-time events for reference and contextualization
- Real-time process can act as a data source with the help of real-time queries

Orchestration

- Model-based orchestration of real-time application resources is possible on DataSwarm platform
- Batch job(s) can be executed in response to real-time events either received or derived by analytics
- Whenever source repositories are updated, platform can dynamically update in-memory reference data
- Preemptive tuning of the application can be done for outage avoidance using predictive analytics (machine learning) for real-time and batch performance and resource use

Support top IoT protocols out of the box

- Out-of-the-box components include:
    - Real-time sources/protocols such as Kafka, MQTT, AMQP, XMPP, Kinesis (AWS)
    - ETL functions (transforms, filters, etc.)
    - Analytics such as anomaly detection, flexible windowing, trending, ranking, and various machine-learning algorithms
    - Alerts (single and multi-event based)

Auto Transformation

- Automatically lay out a pattern of connected components to transform data as needed- Sources, Dashboards, and Applications

Meta Composition

- Use applications as virtual components within another application

Management

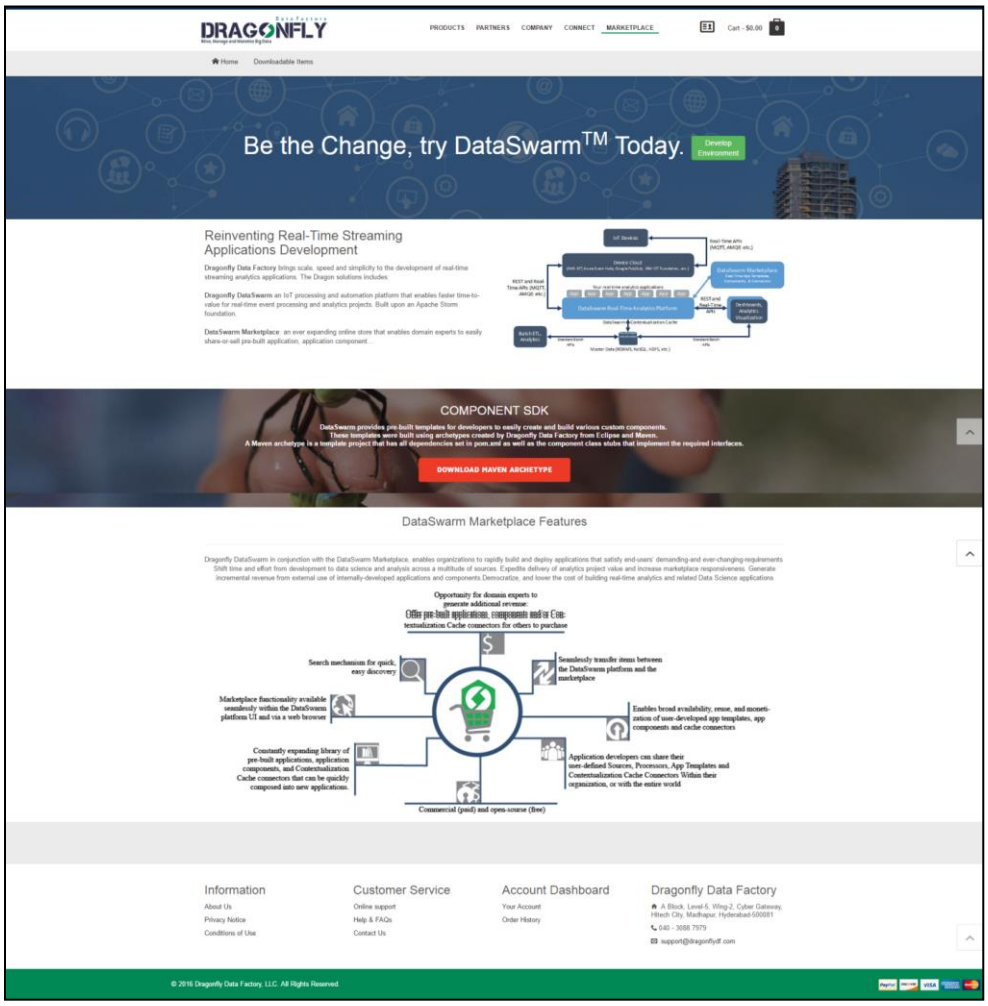
- Integrated management of applications like - Deployment, un-deployment, rebalancing, starting, stopping, monitoring, etc.
- Integrated management of execution-engine clusters. For e.g. Add/remove nodes, performance monitoring, etc.
- Integrated monitoring of data sources and integration agents like - Message brokers, repositories, in-memory DB caches, etc., performance, resource monitoring and configurable alerts
- Integrated infrastructure management of underlying cloud (public, private & hybrid) resources
- Multi-tenant support (secure separation)

DataSwarm Marketplace Overview

The DataSwarm Marketplace (DSM) is a digital distribution platform similar to an app store, for DataSwarm products, developed and maintained by Dragonfly Data Factory. The service allows users to upload, browse and download components, applications and connectors that are developed using Dragonfly's DataSwarm SDKs or platform. The products can be uploaded and downloaded directly from and to the DataSwarm platform.

Marketplace categorizes products based on different verticals or industries. A user can get into specific vertical and browse for relevant components and use case applications for specialized needs, and also can contribute components and applications pertaining to his domain or industry.

Marketplace is typically a form of an online store, where users can browse through these different resource categories, view information about each component or application (such as descriptions, reviews or ratings), and acquire those either by purchase or at no cost. The selected component or application is offered as an automatic download, after which the component or application gets added into the DataSwarm library, simultaneously users can also develop new components and applications and publish those on the Marketplace directly through DataSwarm platform.



DataSwarm Marketplace is curated by Dragonfly Data Factory, requiring that submissions of prospective component and application go through an approval process. These products are inspected for compliance with certain guidelines (such as those for quality control and censorship), including the requirement that a commission be collected on each sale of a paid resource.

DataSwarm Marketplace Features

- DataSwarm Marketplace is an online library of reusable, GUI-configurable components and application templates.
- This global online component library is provided and maintained by Dragonfly and constantly being enhanced and refreshed by Dragonfly and ecosystem of developers
- Marketplace resources can be put as open-source or commercial
- Marketplace has an option to build local online component library for collaboration, sharing and reuse of components within an organization
- Snap-in library for sharing components and templates across the enterprise

Categorization

This marketplace categorizes products based on different verticals or industries. A user can get into specific vertical and browse for relevant components and use-case applications for specialized needs, and also can contribute components and applications pertaining to his domain or industry.

Seamless Transfer

Instead of just download, the marketplace also offers a seamless transfer feature.

Seamless transfer is a process of transferring products from one platform to another without any intervention. Here in this case the products are transferred between DataSwarm platform and DataSwarm Marketplace.

Software Development Kit (SDK)

The SDKs are available in Marketplace for the users, to build components or connectors. These SDKs can be downloaded from the Marketplace, which would enable the users to build products as per the DataSwarm configurations. Different SDKs compliant to different programing languages like Java, SQL etc. are available, so the users can choose any SDK as per their requirements.