Ecosystem

Here is a list of tools we have been told about that integrate with Kafka outside the main distribution. We haven't tried them all, so they may not work!

Clients, of course, are listed separately here.

Kafka Connect

Kafka has a built-in framework called Kafka Connect for writing sources and sinks that either continuously ingest data into Kafka or continuously ingest data in Kafka into external systems. The connectors themselves for different applications or data systems are federated and maintained separately from the main code base. An externally hosted list of connectors is maintained by Confluent at the Confluent Hub.

Distributions & Packaging

- Confluent: Confluent Cloud (SaaS, fully managed) and Confluent Platform (software download for self-managed/on-premise)
- Cloudera distribution - https://www.cloudera.com/products/open-source/apache-hadoop/apache-kafka.html
- EmblocSoft distribution and CPFA training - https://www.emblocsoft.com/pages/en/solutions/kafka/support/
- IBM Event Streams - https://www.ibm.com/cloud/event-streams - Apache Kafka on premise and the public cloud
- Strimzi - http://strimzi.io/ - Apache Kafka Operator for Kubernetes and Openshift. Downloads and Helm Chart - https://github.com/strimzi/strimzi-kafka-operator/releases/latest
- TIBCO Messaging - Apache Kafka Distribution - https://www.tibco.com/products/apache-kafka Downloads - https://www.tibco.com/products/tibco-messaging/downloads

Stream Processing

- Kafka Streams - the built-in stream processing library of the Apache Kafka project
  - Documentation in Apache Kafka
  - Documentation in Confluent Platform
  - Kafka Streams code examples in Apache Kafka
  - Kafka Streams code examples provided by Confluent
- Kafka Streams Ecosystem:
  - Complex Event Processing (CEP): https://github.com/fhussonnois/kafkastreams-cep.
  - Fluent Kafka Streams Test: https://github.com/bakdata/fluent-kafka-streams-tests (blog post: https://medium.com/bakdata/fluent-kafka-streams-tests-e641785171ec)
  - Azkarra Streams - A lightweight java framework to make it easy to build and manage streaming microservices based on Kafka Streams.
- Storm - A stream-processing framework.
- Samza - A YARN-based stream processing framework.
- Storm Spout - Consume messages from Kafka and emit as Storm tuples
- Kafka-Storm - Kafka 0.8, Storm 0.9, Avro integration
- SparkStreaming - Kafka receiver supports Kafka 0.8 and above
- Flink - Apache Flink has an integration with Kafka
- IBM Streams - A stream processing framework with Kafka source and sink to consume and produce Kafka messages
- Spring Cloud Stream - a framework for building event-driven microservices, Spring Cloud Data Flow - a cloud-native orchestration service for Spring Cloud Stream applications
- Apache Apex - Stream processing framework with connectors for Kafka as source and sink.
- Logstash - Input and Output plugins to enrich events and optionally store in Elasticsearch
- Logagent - Kafka Input and Kafka Output plugins

Hadoop Integration

- Confluent HDFS Connector - A sink connector for the Kafka Connect framework for writing data from Kafka to Hadoop HDFS
- Camus - LinkedIn's Kafka=>HDFS pipeline. This one is used for all data at LinkedIn, and works great.
- Kafka Hadoop Loader A different take on Hadoop loading functionality from what is included in the main distribution.
- Flume - Contains Kafka source (consumer) and sink (producer)
- KaBoom - A high-performance HDFS data loader

Database Integration

- Confluent JDBC Connector - A source connector for the Kafka Connect framework for writing data from RDBMS (e.g. MySQL) to Kafka
- Oracle Golden Gate Connector - Source connector that collects CDC operations via Golden Gate and writes them to Kafka

Search and Query

- Elasticsearch - This project, Kafka Standalone Consumer will read the messages from Kafka, processes and index them in Elasticsearch. There are also several Kafka Connect connectors for Elasticsearch.
- Presto - The Presto Kafka connector allows you to query Kafka in SQL using Presto.
- Hive - Hive SerDe that allows querying Kafka (Avro only for now) using Hive SQL
- OpenMLDB Kafka Connector - This project allows you to define and extract features from data streams using SQL for ML applications.

Management Consoles

- Kafka Manager - A tool for managing Apache Kafka.
- kafkat - Simplified command-line administration for Kafka brokers.
- Kafka Web Console - Displays information about your Kafka cluster including which nodes are up and what topics they host data for.
- Kafka Offset Monitor - Displays the state of all consumers and how far behind the head of the stream they are.
- Capillary – Displays the state and deltas of Kafka-based Apache Storm topologies. Supports Kafka >= 0.8. It also provides an API for fetching this information for monitoring purposes.
- Doctor Kafka - Service for cluster auto healing and workload balancing.
- Cruise Control - Fully automate the dynamic workload rebalance and self-healing of a Kafka cluster.
- Burrow - Monitoring companion that provides consumer lag checking as a service without the need for specifying thresholds.
- Chaperone - An audit system that monitors the completeness and latency of data stream.
- Sematext integration for Kafka monitoring that collects and charts 200+ Kafka metrics
- Xinfra Monitor - A framework that monitors and exposes metrics showing availability and performance of Kafka clusters and mirrored pipelines.

AWS Integration

- Automated AWS deployment
- Kafka -> S3 Mirroring tool from Pinterest.
- Alternative Kafka->S3 Mirroring tool

Logging

- syslog (1M)
  - syslog producer : A producer that supports both raw data and protobuf with meta data for deep analytics usage.
  - syslog-ng (https://syslog-ng.org/) is one of the most widely used open source log collection tools, capable of filtering, classifying, parsing log data and forwarding it to a wide variety of destinations. Kafka is a first-class destination in the syslog-ng tool; details on the integration can be found at https://czanik.blogs.balabit.com/2015/11/kafka-and-syslog-ng/ .
- klogd - A python syslog publisher
- klogd2 - A java syslog publisher
- Tail2Kafka - A simple log tailing utility
- Fluentd plugin - Integration with Fluentd
- Remote log viewer
- Logstash integration - Integration with Logstash and Fluentd
- Syslog Collector written in Go
- Klogger - A simple proxy service for Kafka.
- fuse-kafka: A file system logging agent based on Kafka
- omkafka: Another syslog integration, this one in C and uses librdkafka library
- logkafka - Collect logs and send lines to Apache Kafka
- Filebeat Kafka Module - Collect and ship Kafka logs to Elasticsearch (docs)

Flume - Kafka plugins

- Flume Kafka Plugin - Integration with Flume
- Kafka as a sink and source in Flume - Integration with Flume

Metrics

- Mozilla Metrics Service - A Kafka and Protocol Buffers based metrics and logging system
- Ganglia Integration
- Sematext integration for Kafka monitoring
- Coda Hale Metric Reporter to Kafka
- kafka-dropwizard-reporter - Register built-in Kafka client and stream metrics to Dropwizard Metrics
- Metricbeat Kafka Module - Capture and ship Kafka *consumergroup* and *partition* metrics to Elasticsearch (docs)

Packing and Deployment

- RPM packaging
- Debian packaginghttps://github.com/tomdz/kafka-deb-packaging
- Puppet Integration
  - https://github.com/miguno/puppet-kafka
  - https://github.com/whisklabs/puppet-kafka
- Dropwizard packaging

Kafka Camel Integration

- https://github.com/ipolyzos/camel-kafka
- https://github.com/BreizhBeans/camel-kafka

Misc.

- Kafka Websocket - A proxy that interoperates with websockets for delivering Kafka data to browsers.
- KafkaCat - A native, command line producer and consumer.
- Kafka Mirror - An alternative to the built-in mirroring tool
- Ruby Demo App

- Apache Camel Integration
- Infobright integration
- Riemann Consumer of Metrics
- stormkafkamom – curses-based tool which displays state of Apache Storm based Kafka consumers (Kafka 0.7 only).
- uReplicator - Provides the ability to replicate across Kafka clusters in other data centers
- Mirus - A tool for distributed, high-volume replication between Apache Kafka clusters based on Kafka Connect
- libbeat - All Elastic Beats (Metricbeat, Filebeat, etc) have Kafka outputs