CLOUDERA

Cloudera DataFlow - Virtual Hands-on Workshop

October, 2021

Cloudera Team

Suresh MR
Director - Channels & Alliances

Vinay Rayker
Partner Technology Leader

Puneet Joshi
Partner Solutions Engineer

Pannag Katti
Partner Solutions Engineer

Manick Mehra
Partner Solutions Engineer

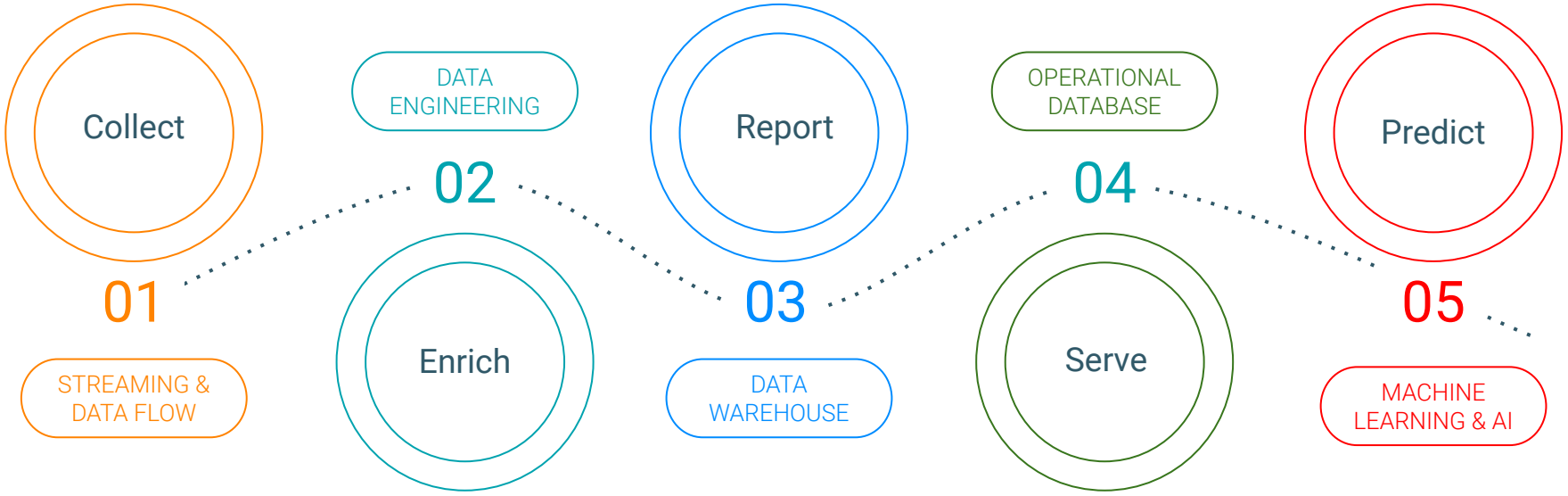
AGENDA

- Cloudera DataFlow - Overview (20 minutes)
- Workshop Overview & Detailed Instructions (10 minutes)
- Hands-on exercises (120 minutes)
- Q&A (throughout the workshop)

Cloudera Data Flow

CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter

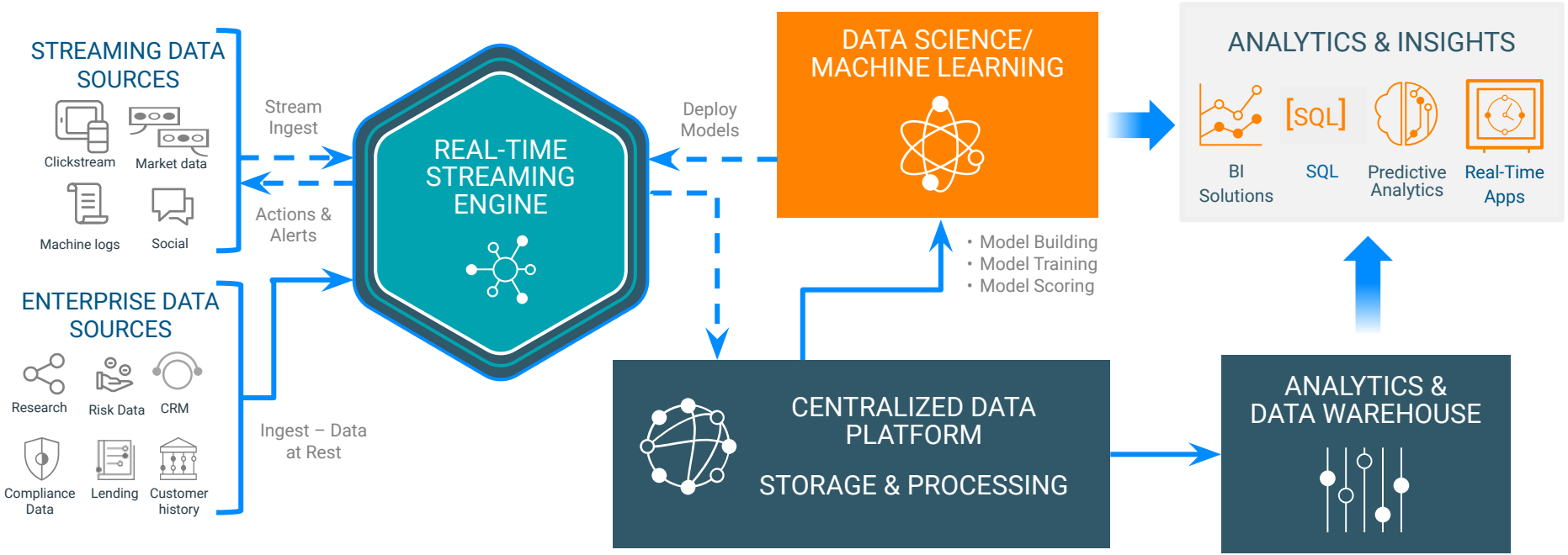


CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

ENABLING ANALYTICS AND INSIGHTS ANYWHERE

Driving enterprise business value



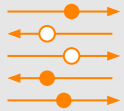
Cloudera Data Flow

REAL-TIME STREAMING ENGINE

Edge Management

Edge data collection, Routing and monitoring

- MiNiFi
- Edge Flow Manager
- NiFi Registry



Flow Management

Enterprise data ingestion, transformation and enrichment

- Apache NiFi
- NiFi Registry

Stream Processing

Real-time stream processing at IoT scale

- Apache Kafka
- Schema Registry

Streams Management

- Streams Messaging Manager
- Streams Replication Manager

Streaming Analytics

Predictive analytics and real-time insights

- Kafka Streams
- Apache Flink
- Spark Streaming

Enterprise Services

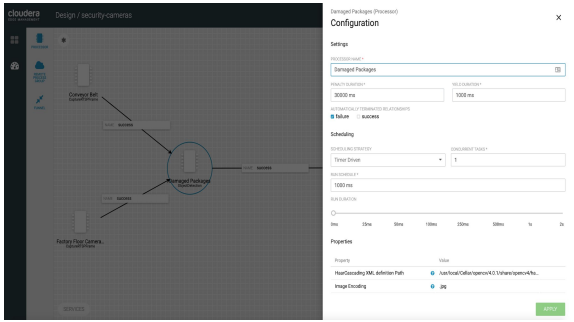
Provisioning, Management and Monitoring

- Unified Security
- Edge-to-Enterprise Governance
- Single Sign-On

Cloudera EDGE Management

Edge device data collection and processing with easy to use central command and control

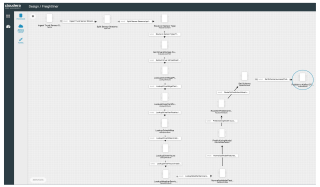
Edge Flow Manager



APACHE NIFI minifi

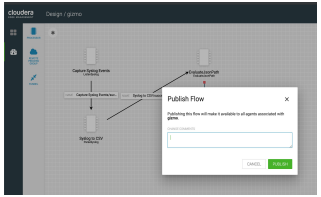
A lightweight edge agent that implements the core features of Apache NiFi, focusing on data collection and processing at the edge

Flow Authorship



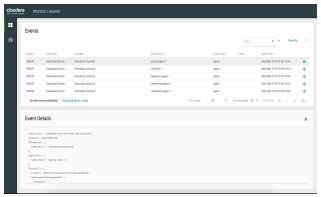
- Small footprint agent with MiNiFi
- Java and C++ agents
- Rich edge processors (edge collection & processing)
- End to end lineage and security

Flow Deployment



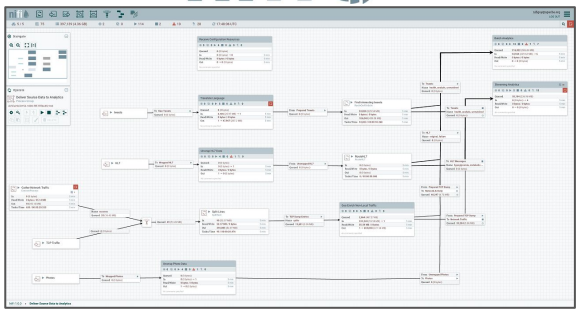
- Central Command and Control (C2)
- Design and deploy to thousands of agents
- Edge Applications lifecycle management
- Multitenancy with Agent classes
- Native integration with other CDF services

Flow Monitoring



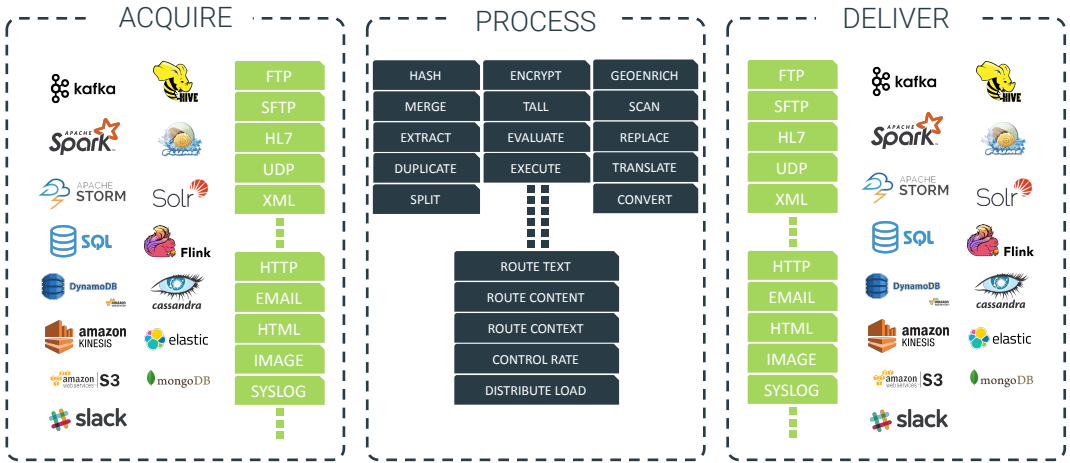
Cloudera Flow management

Enable easy ingestion, routing, management and delivery of any data anywhere (Edge, cloud, data center) to any downstream system with built in end-to-end security and provenance



APACHE NIFI registry

Advanced tooling to industrialize flow development (Flow Development Life Cycle)



- Over 300 Prebuilt Processors
- Easy to build your own
- Parse, Enrich & Apply Schema
- Filter, Split, Merger & Route
- Throttle & Backpressure
- Guaranteed Delivery
- Full data provenance from acquisition to delivery
- Diverse, Non-Traditional Sources
- Eco-system integration

Stream processing & Analytics

Providing simpler ways of complex event processing at scale for the enterprise

kafka + KStreams

"Pub & Sub"

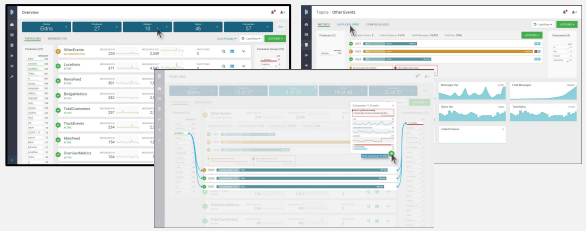
Massively Scalable Publish -Subscribe Message Queue, API for building real-time microservices with Kafka Streams

"Process"

Reliable real-time/micro-batch data processing engines able to process millions tuples/second/node

Streams Messaging Manager

"Manage & Monitor"



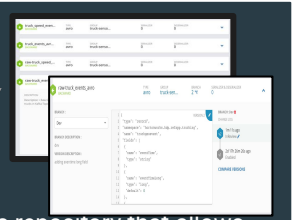
- Cure Kafka blindness and help the different streaming personas be more productive
- End-to-end integration with Ambari, Grafana, Ranger & Atlas
- Comprehensive REST service for open integration

Streams Replication Manager (SRM)

"Mirroring & DR"

Replicate data & configuration in Realtime, Smart clients for easy fail-over & fail-back, Support active active scenarios, Monitoring

SCHEMA REGISTRY



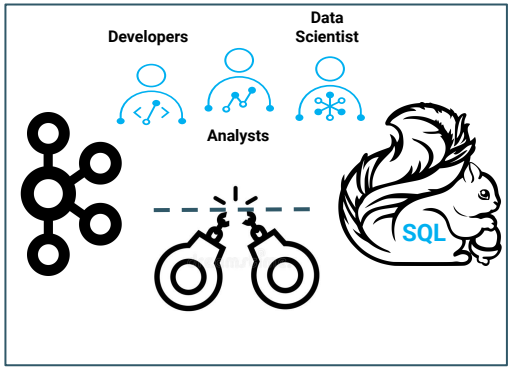
"Manage"

A shared schemas repository that allows applications to save, retrieve and reuse schemas and interact with each other

CLOUDERA STREAMING ANALYTICS INCL. SQL STREAM BUILDER

Democratize access to real-time data with just SQL

Runs on Apache Flink



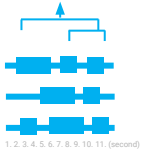
Liberates access to data within Kafka and Flink

No need for Java and Scala experts

Enabling Streaming SQL (aka Continuous SQL)

Query language for processing live data streams

Event Processing

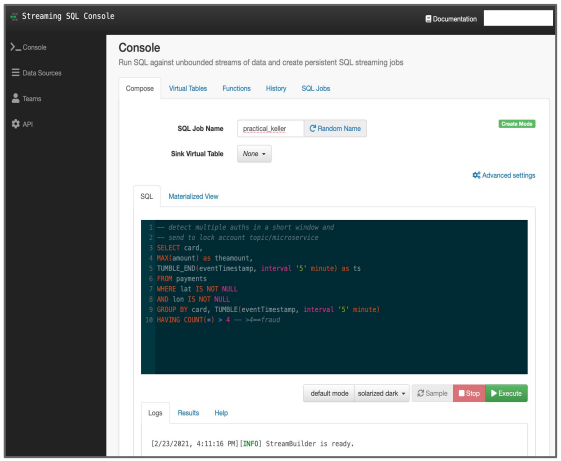


Capture events that matter

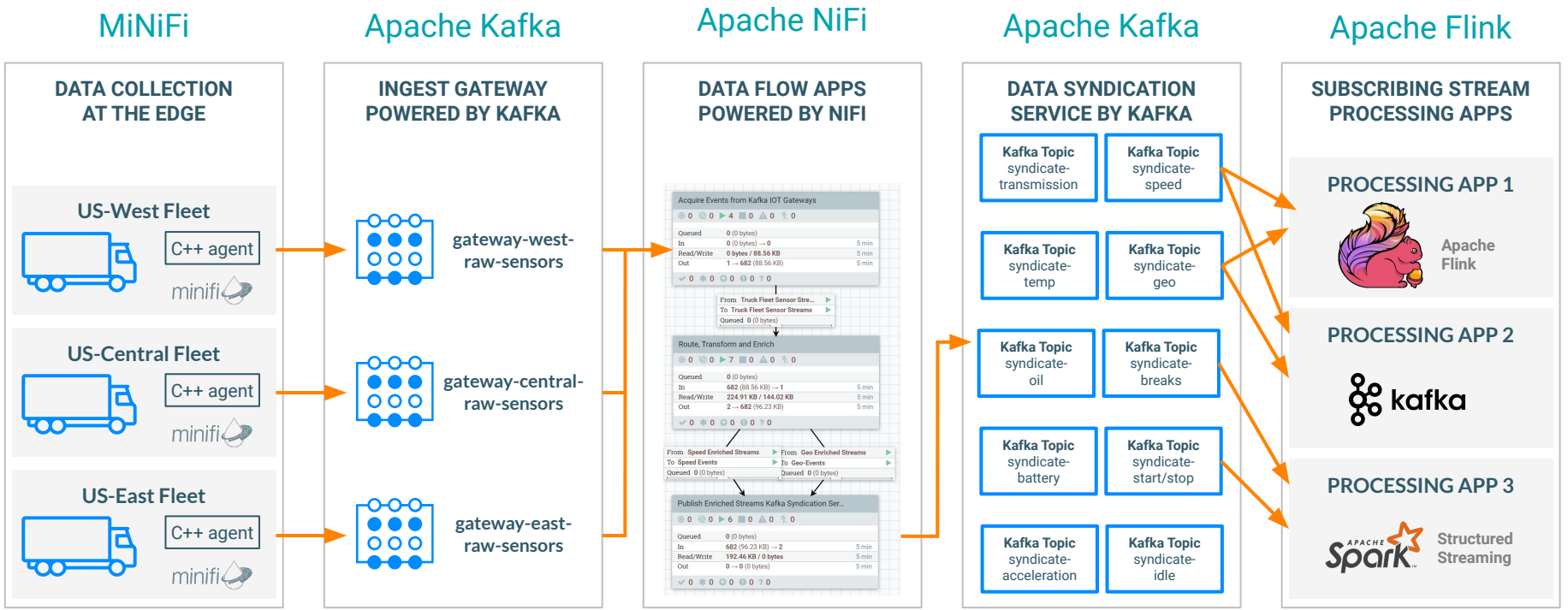
Parsing and blending of offline and streaming data

Streaming Analytics across the line of businesses

Streaming SQL Console



Data-in-Motion Reference Architecture



Workshop Overview

Predicting machine failures

Photos by Dominik Vanyi on Unsplash

Sensors

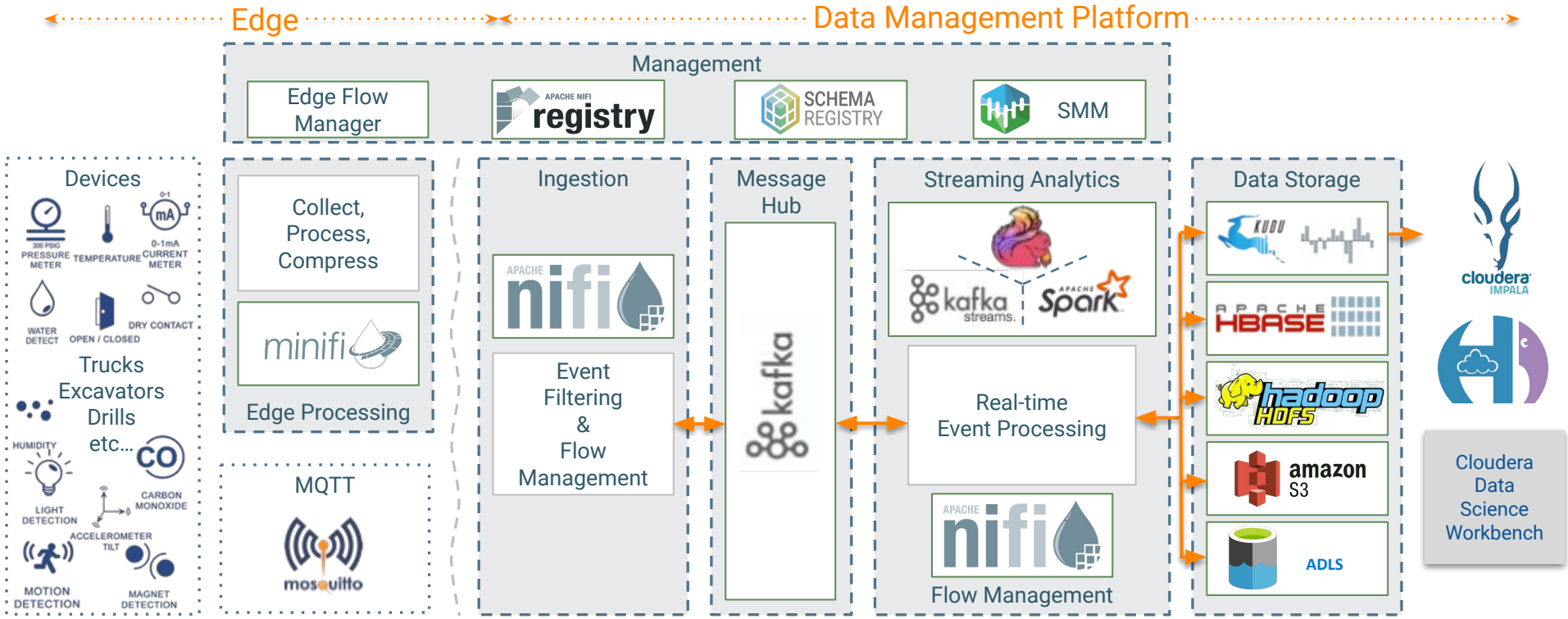
Photo by Kai Dahms on Unsplash

- Temperatures of different machines
- 12 per plant
- Data is published in {"JSON": "format"} to a MQTT server

Streaming and Analytics

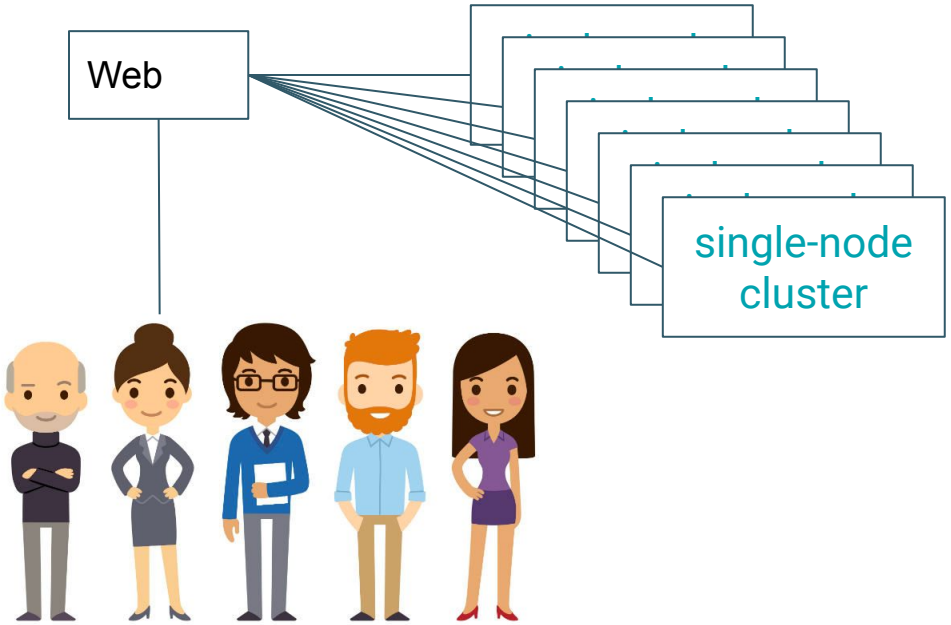
- Process the real-time data and it's collected from plants sensors
- Collect data in our central NiFi, update and forward it to our main Kafka message bus
- Read sensor readings from Kafka and persist into Kudu
- Enable fast analytics on Kudu
- Deploy a small Streaming Analytics application based on Apache Flink which read data stream from kafka topic, summarizing it and publish result into a new topic

And that's what we are doing today



The demo environment

Each participant receives a single-node cluster



Hardware:

- Instance Size m5.4xlarge
- vCPU 16
- Memory (GiB) 64

Software:

- OS CentOS Linux 7.6
- Cloudera **CDP PvC Base 7.1.6**
  - Flow Management 2.0
  - Stream Processing 3.0

How do I get access to a cluster?