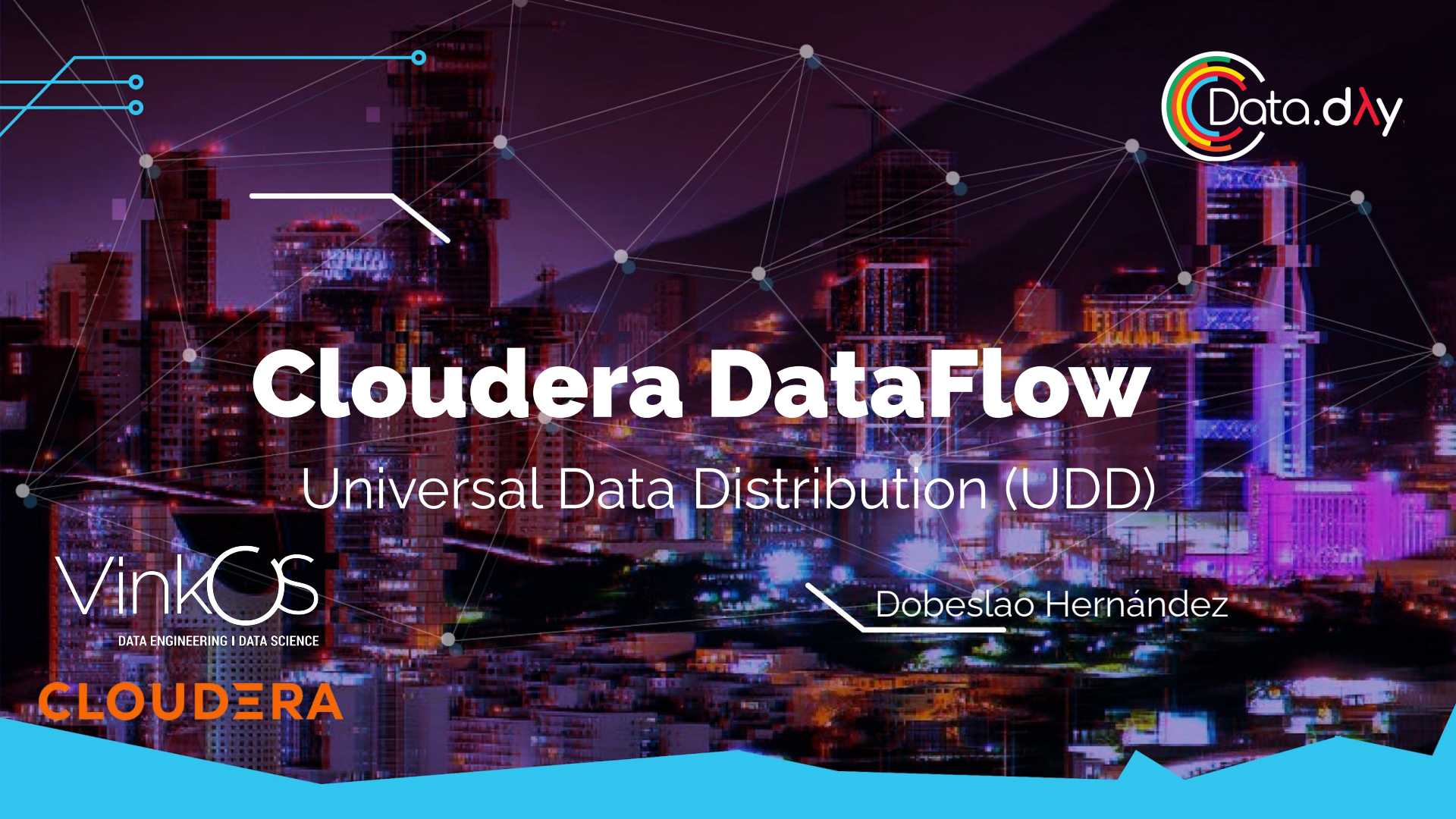


Cloudera DataFlow

Universal Data Distribution (UDD)

Dobeslao Hernández

VinkOS
DATA ENGINEERING I DATA SCIENCE

CLOUDERA

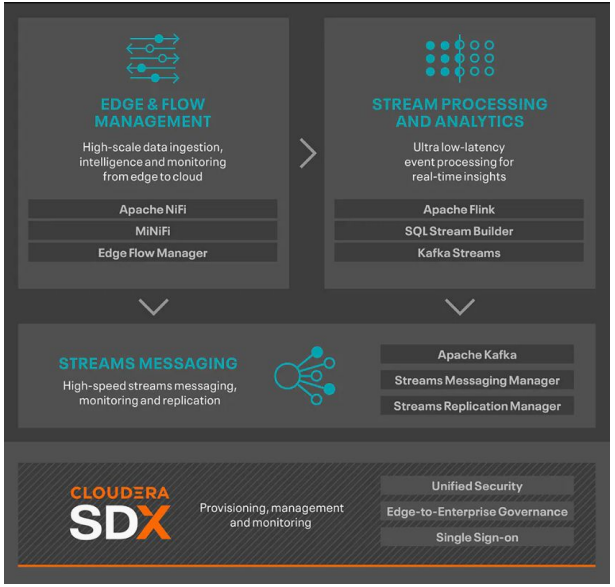
Agenda

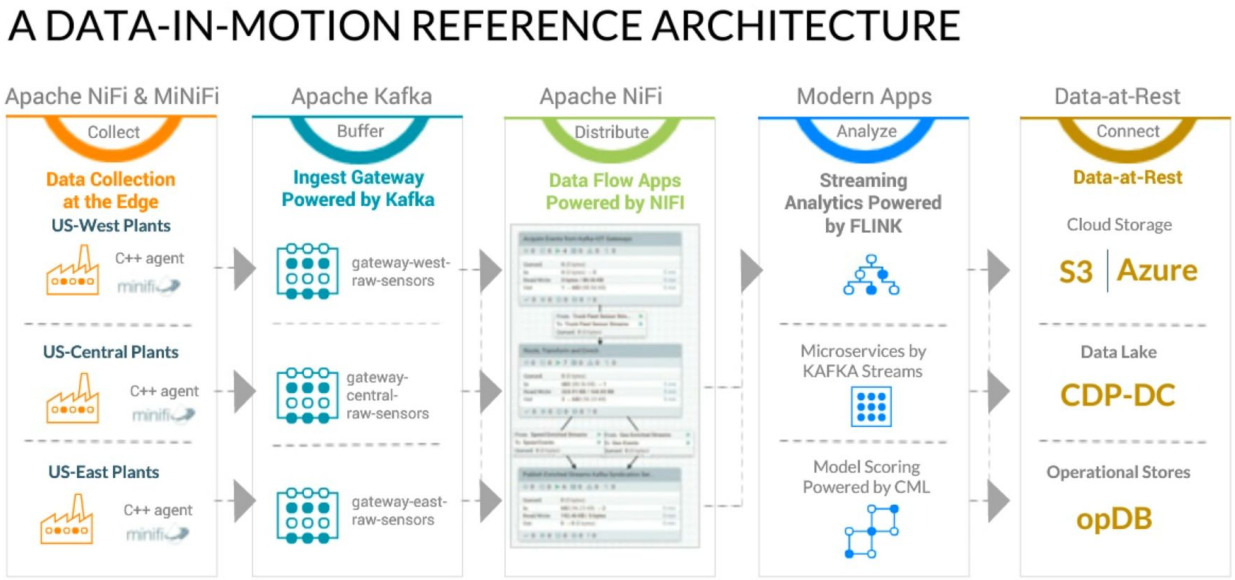
Universal Data Distribution

1. Cloudera DataFlow
2. Componentes
3. Arquitecturas Modernas
4. Demostración

Cloudera DataFlow

La familia de productos Cloudera DataFlow soporta procesos de ingesta continua, intercambio de información entre aplicaciones con todos los elementos de una plataforma empresarial: seguridad, gobierno, auditoría, linaje…





Cloudera DataFlow - Casos de Uso

Algunos casos de Uso …

- **Actividad de sitios web:** seguimiento de visitas, búsquedas, etc. en tiempo real
- **Agregación de eventos y registros:** sobre sistemas distribuidos, mensajes de múltiples fuentes
- **Data Movement:** Optimización de uso de recursos en movimientos de datos entre data centers o entre infraestructura on-premise y nube.
- **Stream Processing:** Combinar múltiples flujos de datos en tiempo real, enriquecer los datos y enviarlos a diferentes destinos en función de reglas definidas
- **Streaming Analytics:** Análisis de patrones en flujos de datos utilizando modelos de machine learning, generando inteligencia accionable.
- **Single view / 360° view of customer**: Ingestar, transformar y combinar datos del cliente provenientes de múltiples fuentes en un punto único de verdad, data lake.
- **Optimize Log Collection & Analysis**: Optimización de soluciones de análisis de logs utilizando Data Flow como la plataforma única para recolectar y entregar múltiples fuentes de datos
- **Capture IoT Data**: Ingestar datos provenientes de sensores de dispositivos IoT y enviarlas para procesamiento y análisis posterior

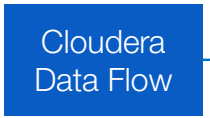
Agenda

Universal Data Distribution

1. Cloudera DataFlow
2. **Componentes**
3. Arquitecturas Modernas
4. Demostración

Cloudera DataFlow - Componentes

Los componentes de Cloudera DataFlow incluyen características empresariales para el desarrollo de soluciones …



Kafka

- Sistema de mensajes distribuido, escalable, se basa en publicación y subscripción de mensajes (tópicos).
- Utiliza principalmente el modelo push para ingestar datos.
- Desacopla aplicaciones, permite patrones de “muchos a muchos”
- Organiza por tópicos para soportar múltiples casos de uso
- Operación sobre ventanas de tiempo
- Integra herramienta de monitoreo, notificación y visualización de los tópicos. CDF SMM

Nifi

- Plataforma para desarrollo de flujos automatizados, escalables y administrados entre sistemas
- Permite Recolectar, Preparar, Analizar y Actuar sobre los datos
- Recibe flujos de datos continuos, con seguridad y capacidad de priorización. Sin necesidad de código.
- Permite el desarrollo y la implementación de tipo DevOps con el soporte de NiFi Registry

Flink

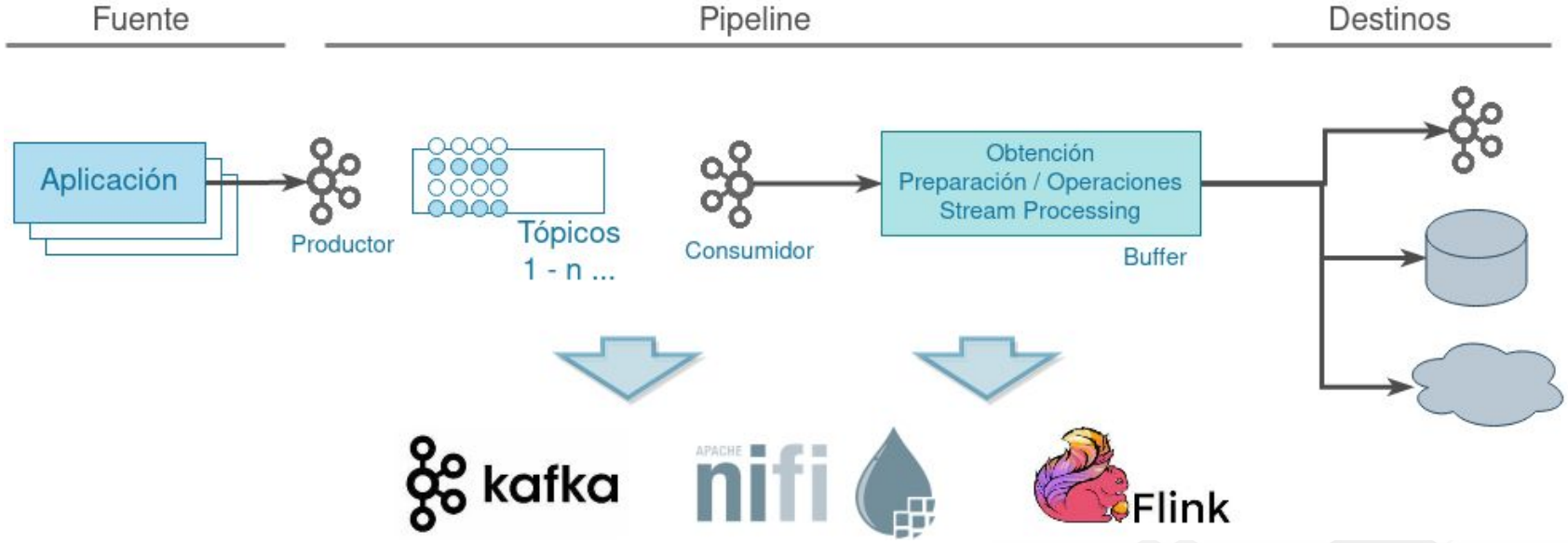
- Sistema distribuido para procesamiento de datos
- Diseñado para hacer analytics en streaming y batch
- Enfoque en aplicaciones basadas en eventos con consultas dinámicas
- Procesamiento en memoria a gran escala / sobre ventanas de t.
- Soporte para SQL + herramientas empresariales CDF SQL Stream Builder

Edge

- Agente para recolectar datos en el origen.
- Componente ligero y poco invasivo en las aplicaciones origen
- Administración centralizada de agentes
- Generación de “data provenance”
- Integración con NiFi

Procesos de Ingesta Continua - Arquitectura de caso de uso implementado

Si trasladamos los puntos anteriores a un proceso de datos, lo visualizamos:



En este tipo de procesos, los datos están en flujo constante: Data-in-Motion

CDF SQL
Stream Builder

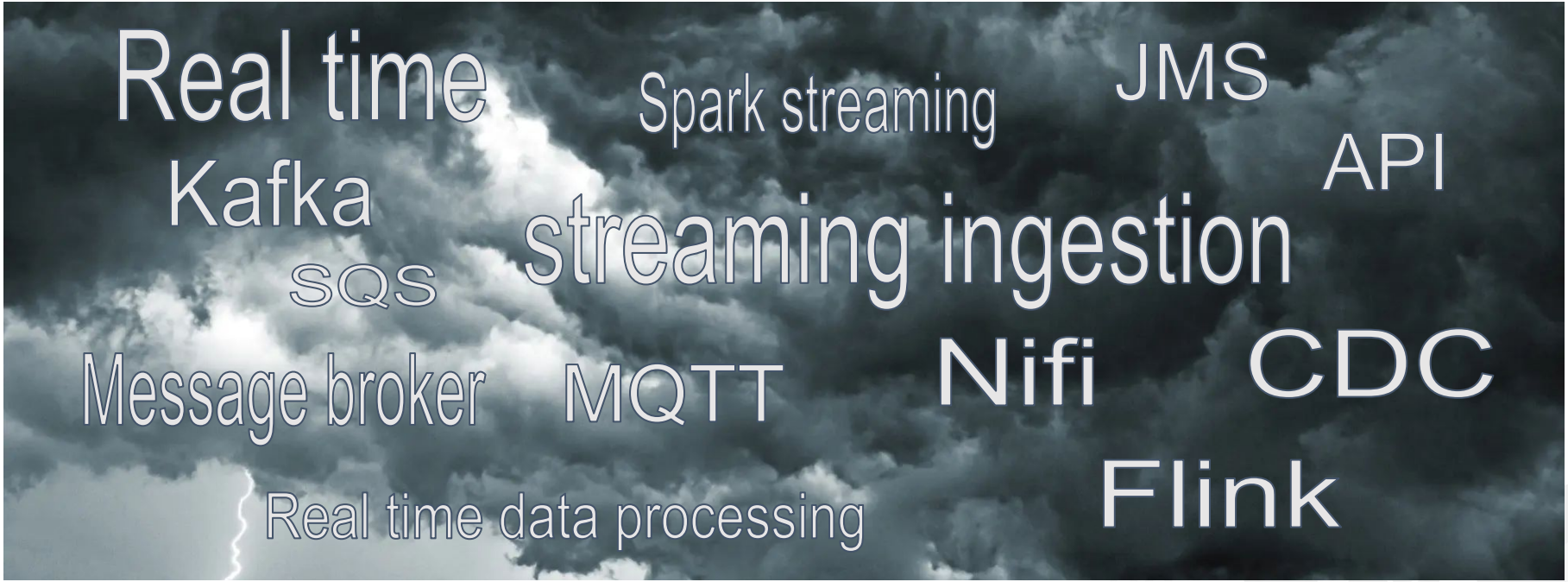
Agenda

Universal Data Distribution

1. Cloudera DataFlow
2. Componentes
3. **Arquitecturas Modernas**
4. Demostración

Procesos de Ingesta Continua …

¿De qué hablamos cuando decimos Ingesta Continua?



… Pero, ¿Cuáles son los elementos, los componentes que debo considerar para empezar con procesos de ingesta continua?

Procesos de Ingesta Continua - Consideraciones

¿Qué necesitamos saber para implementar un proceso de ingesta continua?

Del entorno

- Tipos de fuentes y destinos en mi arquitectura: APIs, Servicios, Transacciones, Logs
- Frecuencia de generación de registros continua o en microbatch

De los componentes

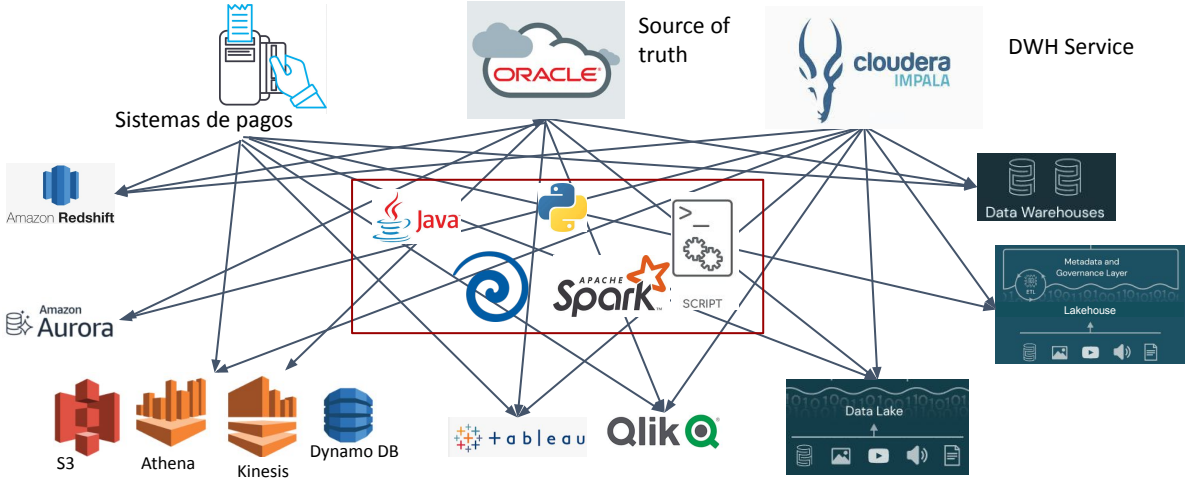
Componentes principales:
- Transporte: Movimiento de datos de una fuente a un destino
- Retención: Capacidad de retener registros en caso de contingencia para no perder información
- Preparación / Enriquecimiento: Capacidad para realizar operaciones sobre el flujo de datos, antes de persistirlos.

De las herramientas

Características:
- Despliegue on premise / en nube
- Desarrollo escribiendo código (java/python, otros)
- Programación por objetos / Parametrización
- Uso de lenguaje SQL
- Integración con Seguridad, administración de usuarios
- Gobierno / Lineage / Auditoría

Retos de las Arquitecturas Modernas

Una arquitectura moderna incorpora múltiples aplicaciones y sistemas que necesitan interactuar enviando y recibiendo datos con diferentes formatos, frecuencias, estructuras; unidireccional o bidireccional:



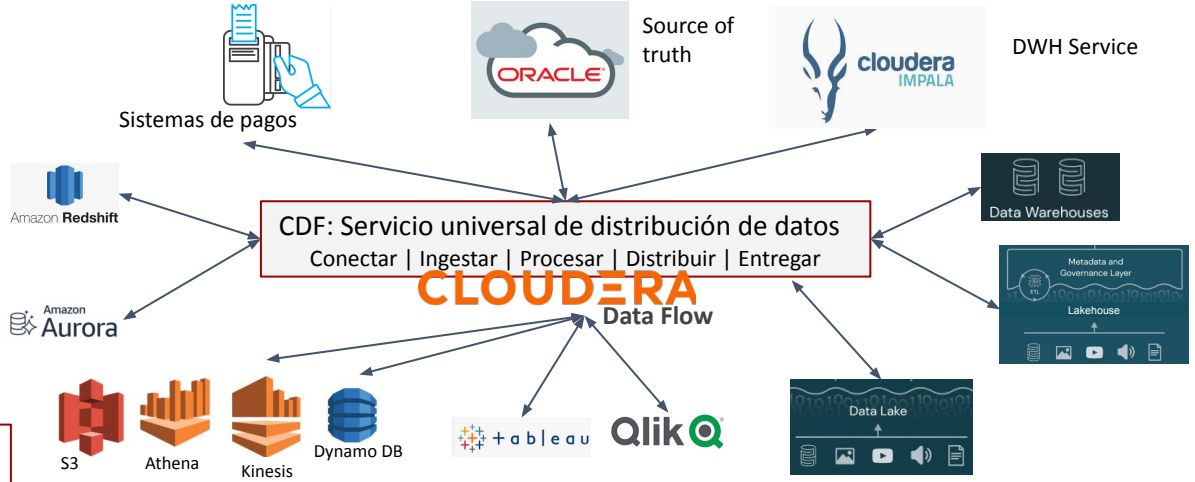
RETOS

- La velocidad para mover los datos eficientemente a través de la Organización .
- Las herramientas de “Extracción/Carga” se enfocan en fuentes on premise o en nube, sin considerar que un entorno de real combina on premise, nube, edge, etc y no necesariamente son estructuradas y con esquemas definidos.
- Se requieren ó se desarrollan manualmente múltiples herramientas para la “Extracción/Carga” para mover los datos a través de los diferentes elementos del ecosistema

Se requiere un servicio universal de distribución de datos

Servicio universal de distribución de datos

Un servicio universal de distribución de datos permite conectar a cualquier fuente de datos, en cualquier lugar, procesar y enviar a cualquier destino:



Capacidades

- Soporta ingesta de datos en entornos híbridos, conectándose a cualquier fuente de datos en cualquier lugar en cualquier nube con cualquier estructura
- No excluye donde distribuye los datos, soporta la entrega a cualquier destino: lakehouses, data lakes, servicios
- Soporta casos de uso diversos: Continuous Data Collection, batch, microservicios, event-driven, streaming…
- Soporta los desafíos de integración de datos: formateo de datos, ruteo, filtrado, administración de errores, reintentos/reejecuciones, interacción con múltiples protocolos.
- El desarrollo es por configuración de objetos, no código; con posibilidad de extensiones para incrementar las funcionalidades

Universal Data Distribution

Universal Data Distribution de Cloudera ofrece la primera solución de ingesta de datos construida para datos híbridos.

El volumen de datos que las empresas capturan y almacenan en su infraestructura on-premise, en la nube y en streaming sigue aumentando.

Permite a las empresas tomar el control de sus flujos de datos, desde el origen hasta todos los puntos de consumo, tanto on-premise como en la nube, de una manera universal que es simple, segura, escalable y rentable. Esto es posible gracias a **Cloudera DataFlow**, la primera solución de ingesta de datos creada para un mundo de datos híbridos. A diferencia de las soluciones basadas en asistentes, específicas del sistema de destino, Cloudera DataFlow proporciona una distribución de datos indiscriminada con **más de 450 conectores y procesadores** a través de un ecosistema de servicios de nube híbrida, incluyendo data lakes, lakehouses, y almacenes de datos en la nube con fuentes de datos locales y edge.

Cloudera DataFlow, es una verdadera solución de ingesta de datos híbridos que aborda toda la diversidad de casos de uso de los datos: por lotes, impulsado por eventos, edge, microservicios y continuo/streaming. Con Cloudera DataFlow, el streaming es tratado como se debería, convirtiendo cualquier fuente de datos en un flujo de datos, soportando la escala de streaming y desbloqueando cientos de miles de clientes generadores de datos.

Agenda

Universal Data Distribution

1. Cloudera DataFlow
2. Componentes
3. Arquitectura Conceptual
4. **Demostración**

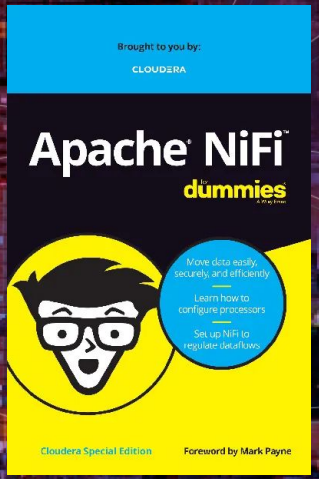

Data.dAy
Demostración
Universal Data Distribution (UDD)
VinkOS
DATA ENGINEERING I DATA SCIENCE
CLOUDERA

Apache® NiFi™ For Dummies®,
Cloudera Special Edition

VinkOS
DATA ENGINEERING I DATA SCIENCE

CLOUDERA

FREE GIFT



Gracias, nos vemos pronto.