ESPECIALIZACIÓN

Ingeniería de datos con Azure

Curso: Streaming Data

Docente: Richard Tadeo Zenteno

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del cha**t y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

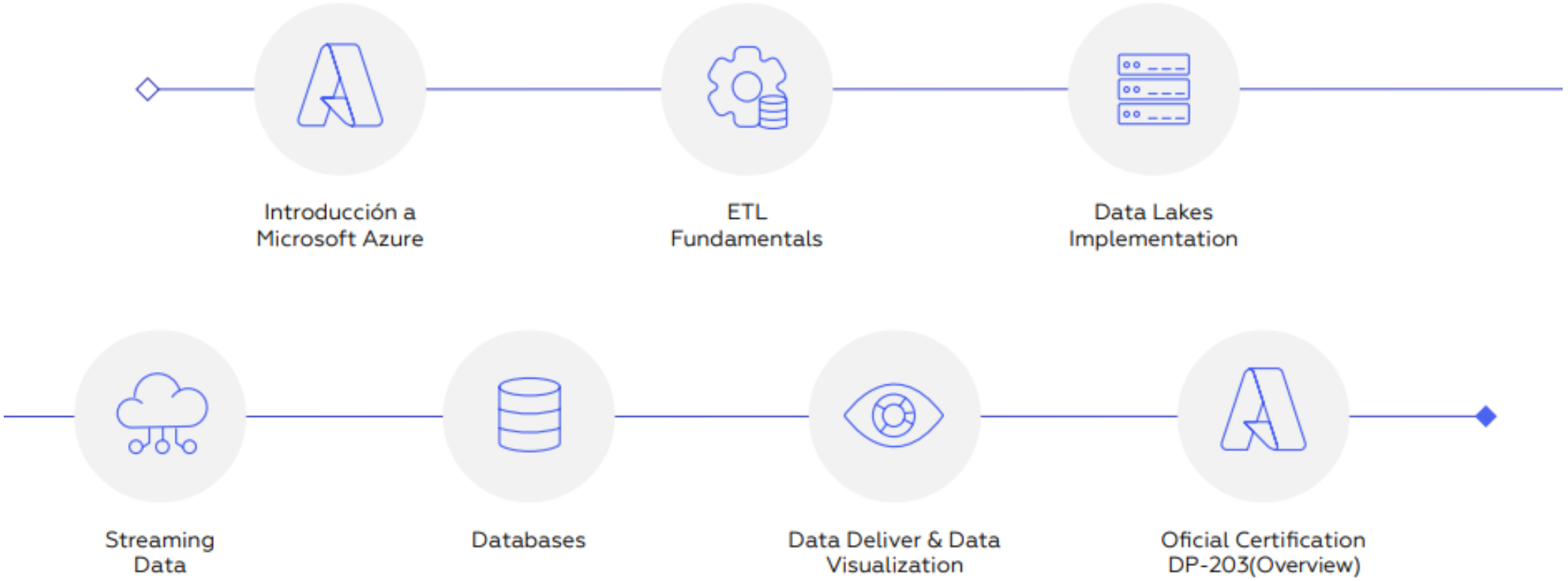
ITINERARIO

07:00 PM – 07:30 PM	***Soporte técnico DMC***
07:30 PM – 08:50 PM	***Agenda***
08:50 PM – 09:00 PM	***Pausa Activa***
09:00 PM – 10:30 PM	***Agenda***

Horario de Atención Área Académica y Soporte

Lunes a Viernes 09:00 am a 10:30 pm / Sábado 09:00 am a 02:00pm

MALLA CURRICULAR



CERIFICACIÓN FINAL
por **Aprobación** de la Especialización en **Ingeniería de Datos con Microsoft Azure** (48 horas académicas)

CONTENIDO

Introducción a Microsoft Azure

- Introducción a Cloud Computing. Proveedores de servicios Cloud, On-Premise vs. On-Cloud, principales servicios, descripción de los modelos de costos.
- Identify and Access Management (IAM). Overview de los roles principales, ejemplos de gestión de permisos.

ETL Fundamentals

- Introducción a las soluciones ETL. Definición, descripción de sus etapas.
- Introducción a los servicios Azure Data Factory y Data Flow. Características generales, casos de uso.
- Taller: Implementación de un ETL Básico con Azure.

Data Lakes Implementation

- Introducción a Data Lakes. Definición, arquitectura, capas (Raw, Stage, Analytics).
- Introducción a los servicios Azure Blob Storage y Storage Account.
- Taller: Implementación de un Datalake en Azure.

CONTENIDO

Streaming Data

- Introducción a procesamiento de datos Batch y Streaming. Diferencias Near-Real-Time y Real-Time.
- Introducción a IoT. Definición, uso de sensores, aplicaciones.
- Revisión de servicios: Azure EventHubs y IoTHub. Características generales, ejemplos de implementación y uso.
- Taller: Manejo de Streaming al Data.

Databases

- Introducción a las bases de datos Relacionales y No-Relacionales. Definición, características, casos de uso.
- Azure SQL Database for MariaDB. Descripción y características generales.
- Azure SQL Database for PostgreSQL. Descripción y características generales.
- Azure SQL Database for CosmosDB. Descripción y características generales.
- Taller: Diseño de una base de datos relacional y técnicas para poblarla.

Data Deliver & Data Visualization

- Azure Synapse Analytics. Propósito del servicio, características generales.
- Fabric. Propósito del servicio, características generales.
- Taller: Conexión de Power BI a servicios de datos de Azure.

AGENDA

01

Introducción a procesamiento de datos Batch y Streaming

02

Revisión de servicios: Azure EventHubs y IoTHub

03

Introducción a IoT

04

Taller: Manejo de Streaming al Data

¿Qué es Streaming?

El streaming es un medio de enviar y recibir datos (como audio y vídeo) en un flujo continuo a través de una red. Esto permite que la reproducción comience mientras se envía el resto de los datos.

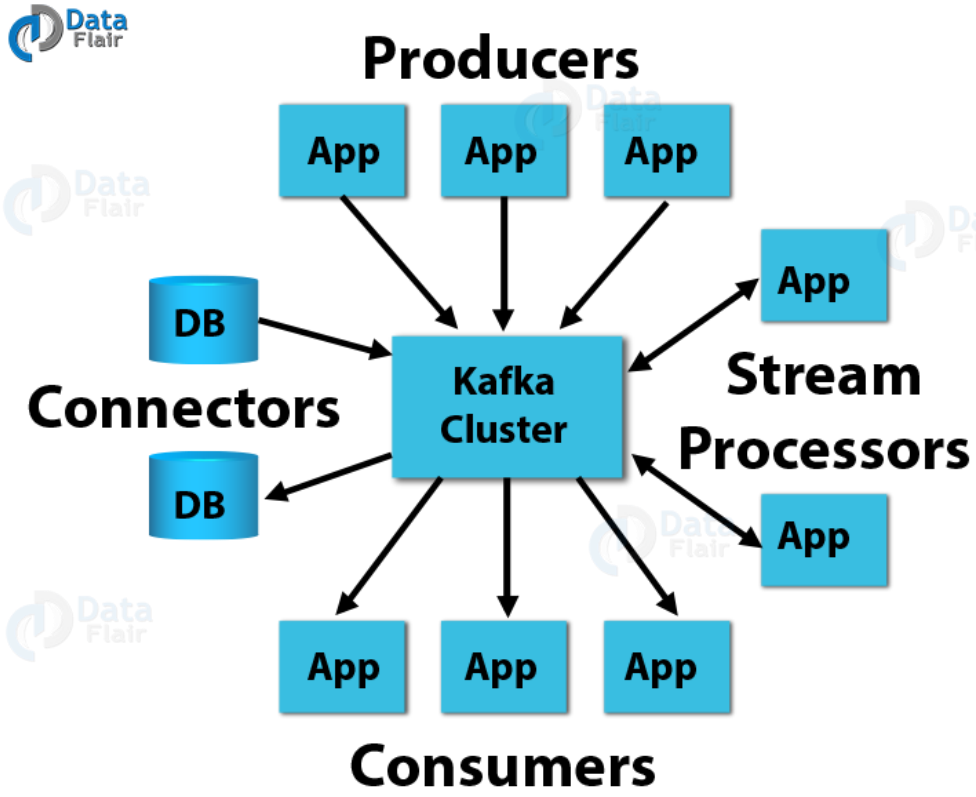
¿Qué es Kafka?

Kafka es un sistema de mensajería **publish-subscribe** que controla las transferencias de datos en forma real-time. Originalmente creado por LinkedIn .

Kafka te ofrece una plataforma para la **construcción de pipelines de datos en real time**. Es altamente escalable y paralelizable. Se ejecuta sobre un clúster.

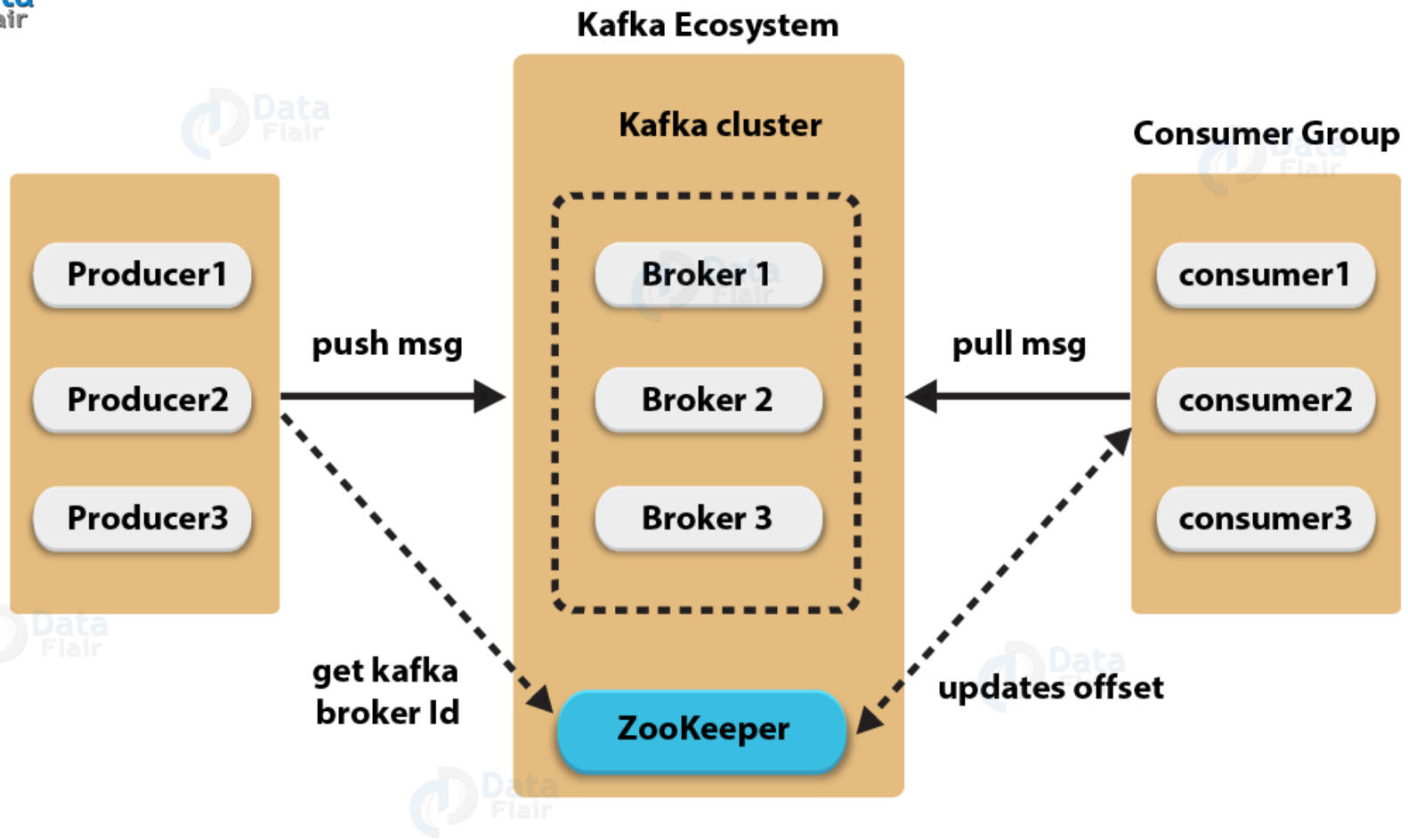
Características de Kafka

- ✓ Escalable
- ✓ Alto rendimiento
- ✓ Soporta ingesta de datos real-time
- ✓ Soporta delivery de datos real-time
- ✓ Flexible
- ✓ Es tolerante a fallos

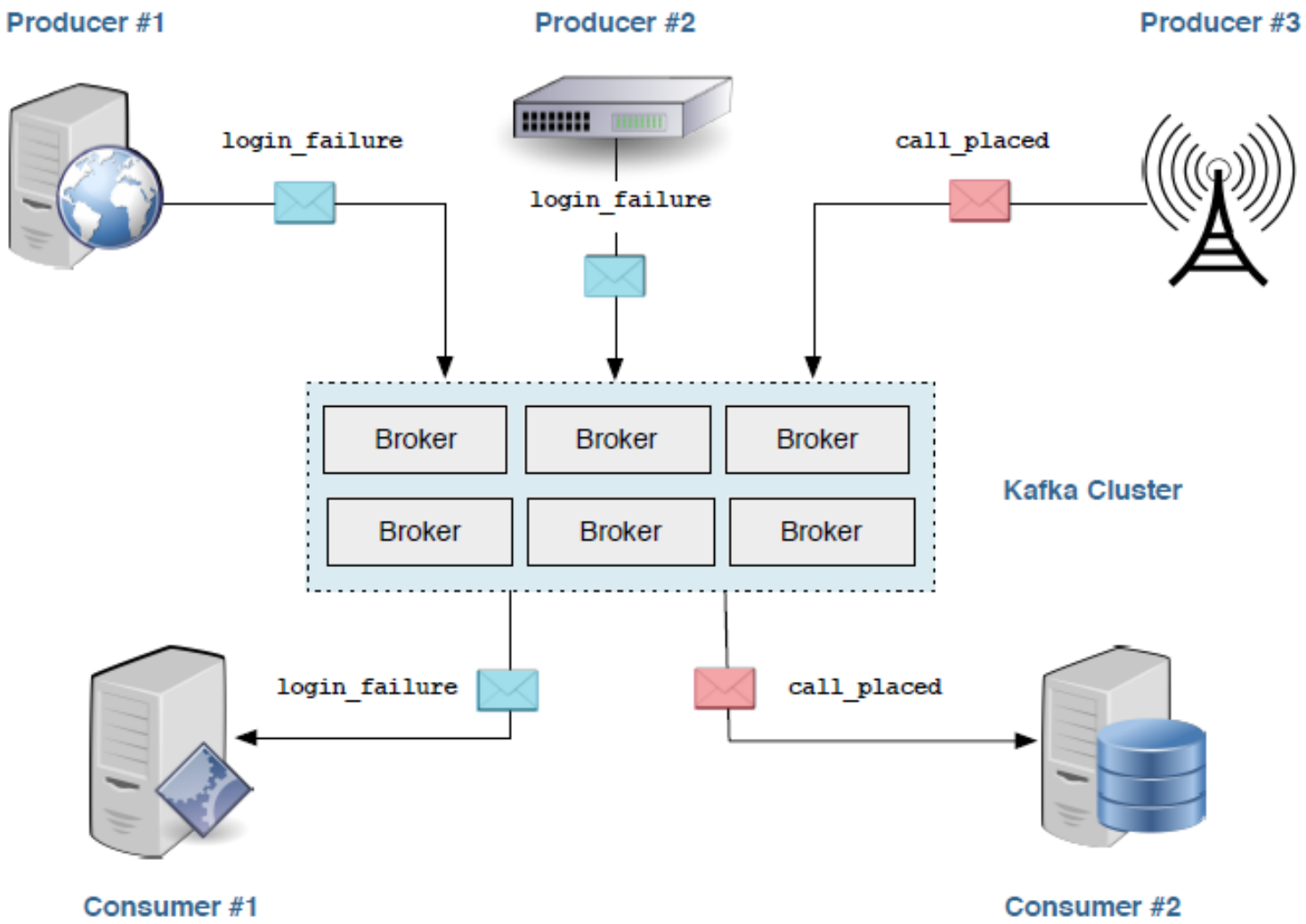


La guía completa de la arquitectura de Kafka (digitalvidya.com)

Arquitectura Kafka



Ejemplo - Arquitectura Kafka

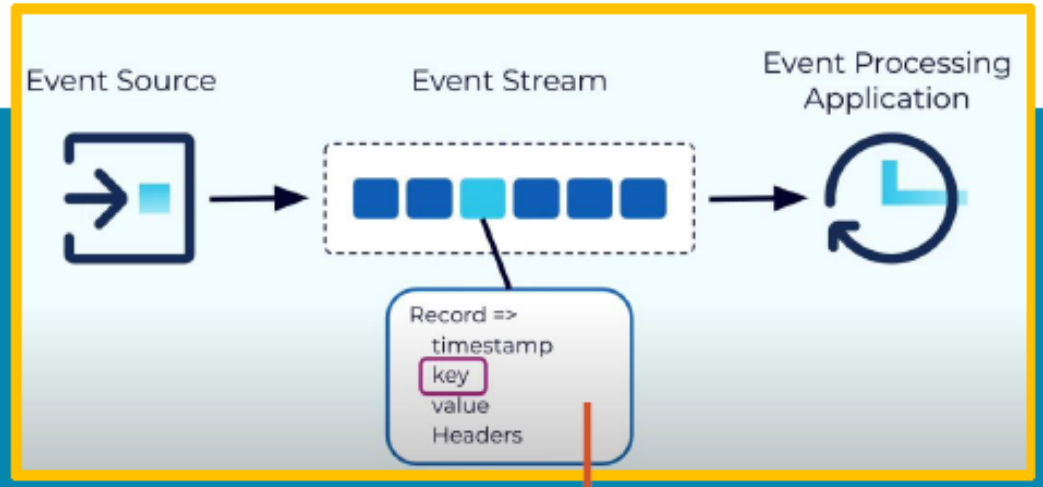


¿Qué son los Eventos?

Hechos ocurridos en las Aplicaciones: Bussiness facts.

Application-level headers

Contiene metadata sobre el Evento



Son immutables

- Pueden ignorarse pero no eliminar, solo un nuevo evento puede invalidar uno anterior.
- Viajan en una sola dirección y no esperan respuesta (fire and forget), pero uno puede "sintetizarse" a partir de un evento posterior.

Keys

Son opcionales, pero con un Rol importante sobre como los eventos son distribuidos a traves de las particiones

Cuando Usarlo:

- Cuando se tiene **varias aplicaciones** que necesitan enterarse del flujo de eventos.
- Cuando se **necesita procesar** los eventos en forma paralela.

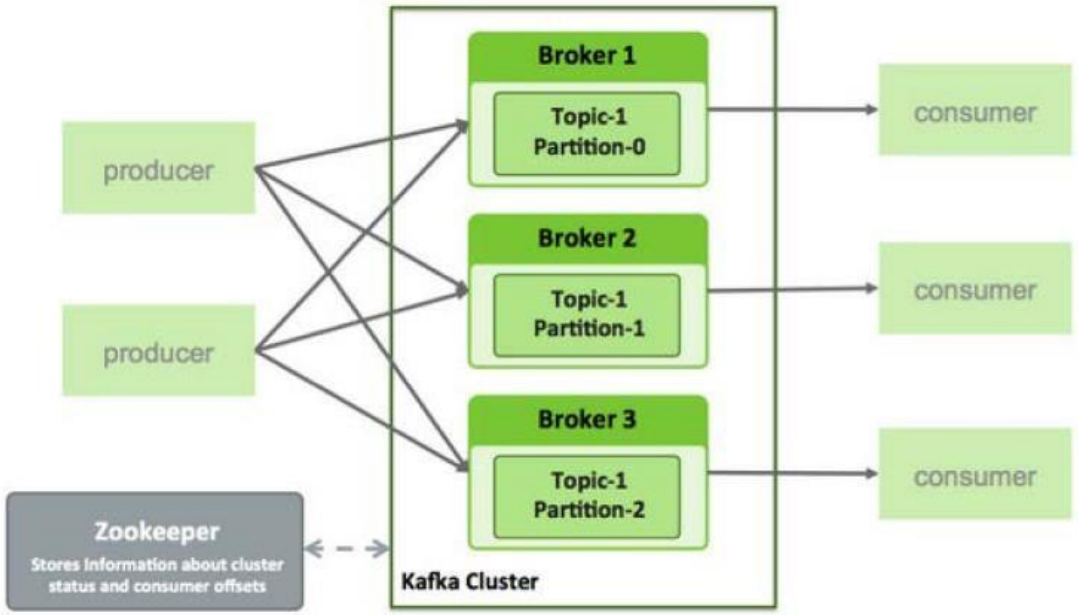
key/ value Bytes	Area	Description
0	Magic Byte	Confluent serialization format version number; currently always 0.
1-4	Schema ID	4-byte schema ID as returned by Schema Registry.
5-...	Data	Serialized data for the specified schema format.

Value

- Representa al contenido **"payload"** del evento.
- Puede estar serializada en **estructuras normalizadas** "AVRO, JSON, Protobuf"

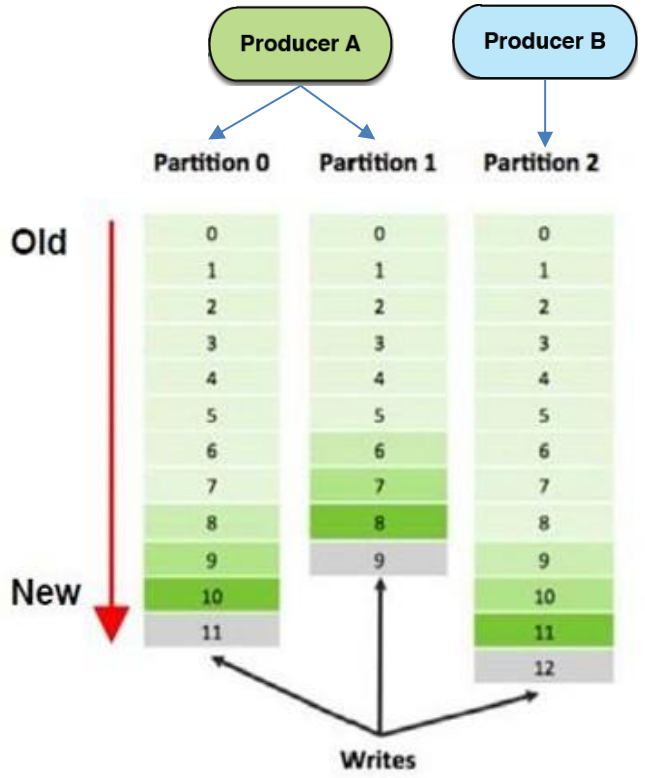
Topics & Partitions

- Los datos se almacenan en Topics los cuales publican los mensajes.
- Topics son divididos en Partitions que son replicados
- En Kafka, los Topics consiste de uno o mós particiones que son secuencias de mensajes que permanecen inmutables y ordenados. Cuando se escribe en una partición estos se almacenan secuencialmente. Este diseño reduce en gran medida el número de lectura a disco.



Topics & Partitions

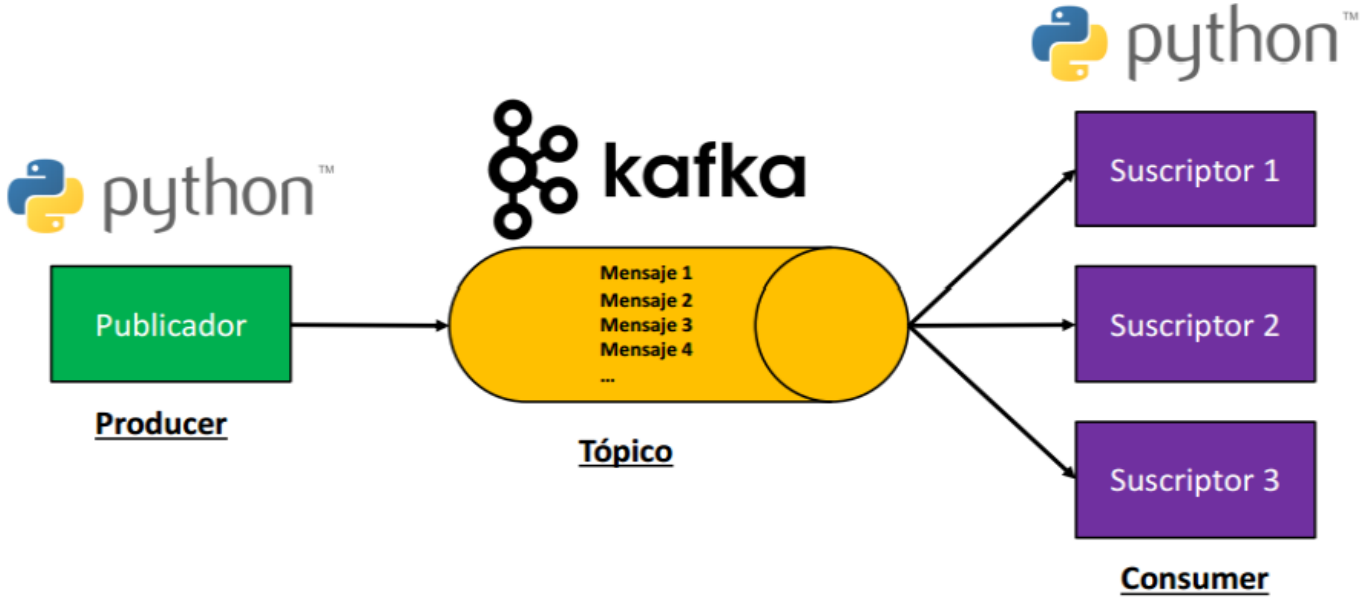
- Cada mensaje en los partitions de los Topics es asignado un secuencia de ID único.
- Después de una cantidad configurable de tiempo (Predeterminado = 7 días), el mensaje publicado se descarta para liberar espacio.
- Los consumers realizan seguimiento de los mensajes que se han consumido.



Kafka en el Cloud

El patrón publicador/suscriptor

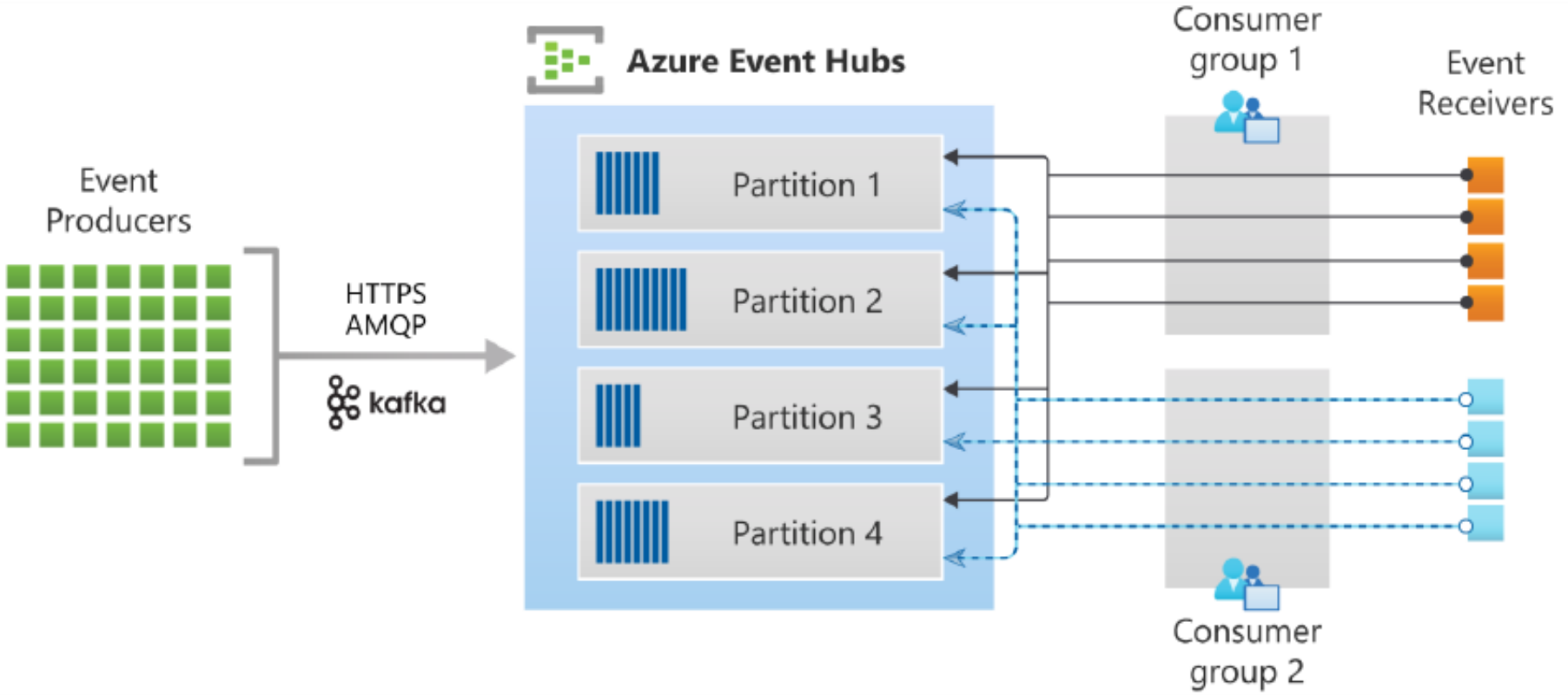
El “producer” y el “consumer” lo puede gestionar Kafka con alguna de sus API, pero generalmente están implementados en algún lenguaje de programación o servicio. El tópico es como una “tabla” que gestiona Kafka.



¿Qué es Azure Event Hub?

Es una plataforma de big data streaming y servicios de ingesta de eventos.

Puede recibir y procesar millones de eventos por segundo.



Azure Event Hub - Componentes

1. **Event Hub** Contenedor lógico para eventos. Agrupa eventos en un hub para su procesamiento.
2. **Consumer Groups** Grupos que permiten que múltiples aplicaciones lean eventos desde el Event Hub de manera independiente.
3. **Producers** Aplicaciones o servicios que envían eventos al Event Hub.
4. **Namespaces** Contenedor que agrupa uno o más Event Hubs.
5. **Event Hub Capture** Función opcional que captura eventos y los almacena en Blob Storage o Data Lake Storage para análisis posterior.

Casos de Uso:
Monitoreo en tiempo real.
Telemetría de aplicaciones.
Análisis de eventos.

Event Producers

Event Hubs Producer

- Los eventos tienen un máximo de 1 MB
- Los eventos enviados pueden especificar una Particion Key, todos aquellos que tengan la misma partition key serán enviados a dicha particion en orden
- Si no se asigna una partition Key se asignaran con Round Robin
- Se debe seleccionar un protocolo de publicación al momento de configurar el producer?

Event publisher identify an a Event Hub

- Shared Access Signature (SAS) token

Ingress

AMQP
Kafka Protocol
HTTPS

Protocolos de publicacion

- AMQP 1.0 : tiene mayores costos de red al inicializar la sesión, tiene un mayor rendimiento
- Kafka 1.0
- HTTPS: requiere una sobrecarga adicional de TLS para cada solicitud.

AMQP: Advanced Message Queuing Protocol

Event data Struct:

- Offset
- Sequence number
- Body
- User properties
- System properties

Nota: Max 1mb por evento

Key architecture components

- Namespace
- Event producer
- Event Hub
- Particions
- Consumer groups
- Event receivers (Consumers)
- Throuputs units

EVENT HUB

Event Hub: El tópico o cola de eventos
Partición: Secuencia de eventos dentro de un Event Hub.

- Un Event Hub puede tener Max 32 partciones
- El Nro. De particiones no pueden modificarse
- Los mensajes se almacenan por orden de llegada
- El nro de particiones en un EH se relaciona directamente con el número d e consumers que espera tener.
- Se recomienda que el nro de particiones en un E.H sea >= que el número Throuputs unit. 1:1

Offsets: es la posición de un evento en una partición, puede ser un valor o un timestamp
Consumer Group: Es una vista de un EventHub permiten que múltiples aplicaciones de consumo tengan una vista separada del flujo de eventos y lean el flujo de forma independiente a su propio ritmo y con sus propios Offsets Solo se accede a particiones a través de un grupo de consumidores (Default) Max 20 x EH estándar tier
Thouputs Units: Unidades de capacidad precompradas
Ingress: hasta 1 MB / seg o 1000 events / seg lo que ocurra primero.
Egress: hasta 2 MB / seg o 4096 events per second.
Automatically scale up Azure Event Hubs throughput units

Consumers

AMQP

Cada partición maneja una sesión AMQP 1.0 que facilita el transporte de eventos segregados por particion

+ Spark

EventHubConsumerClient

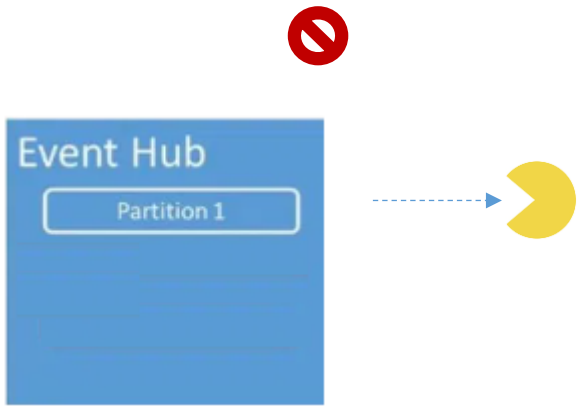
EventConsumer: Entidad que lee datos de un Event Hub

- Se recomienda que solo haya un receptor activo en una partición por grupo de consumidores.
- Un proceso con varios consumer por partición , generara data duplicada, y será necesario manejar esta duplicidad en el código
- Los consumers son responsables de almacenar sus propios valores de Offset fuera del servicio Event Hubs
- Checkpointing: el consumer informa al servicio el offset cuando considere q el flujo de datos esta completo. (Checkponting store BlobStorage)

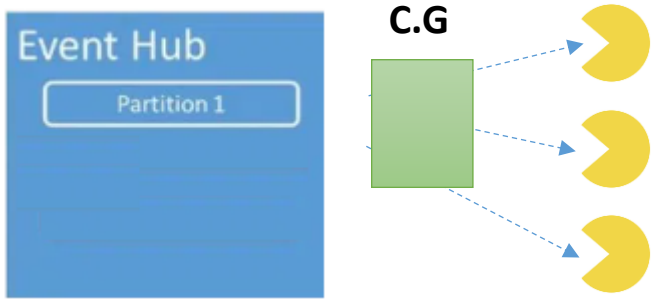
Stream Procesing

APACHE Spark

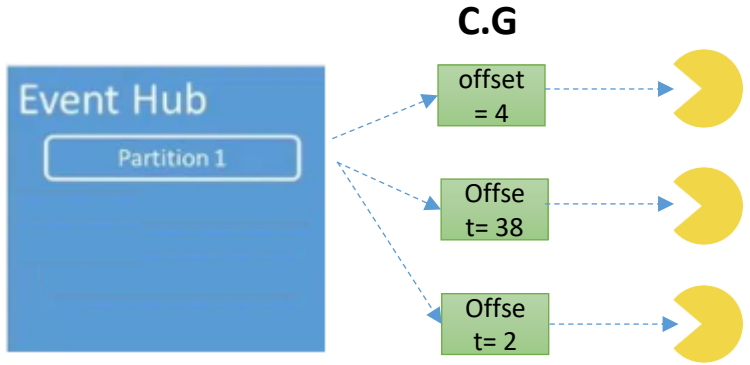
Patrón de consumo en Azure Event Hub



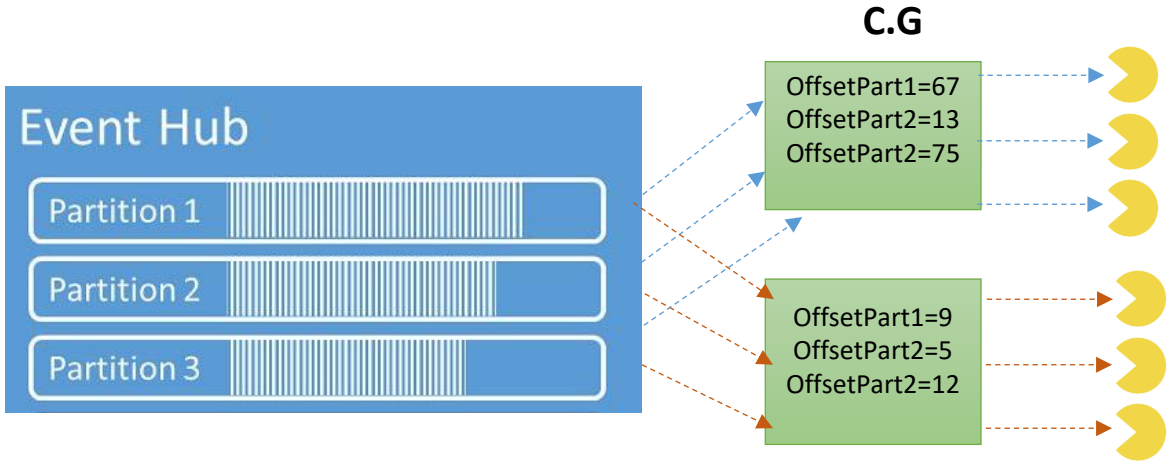
- No ocurre puesto que siempre esta el Default Consumer Grup

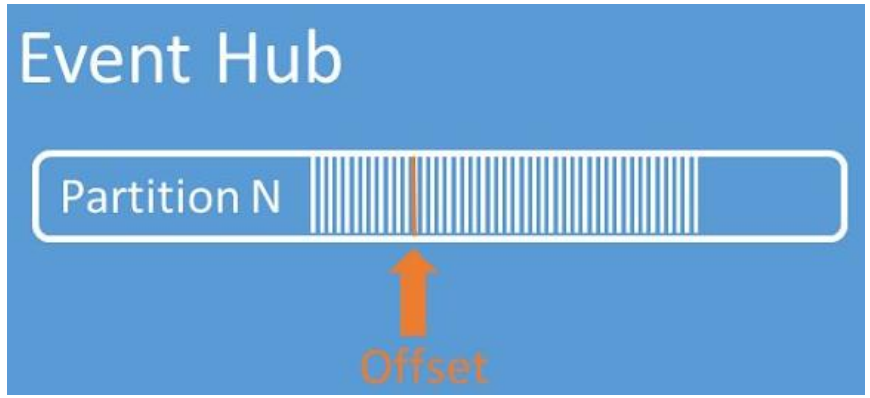


- Los 3 consumer leen data duplicada. No es recomendable mas de un consumer por partición



- Cada consumer a través de su Consumer group tiene el control de lo que ha consumido





¿Qué es Azure IoT Hub?

Es una plataforma de comunicación bidireccional entre dispositivos IoT y la nube. Facilita la gestión, comunicación y supervisión de dispositivos IoT a gran escala, proporcionando una infraestructura segura y escalable.

Cuando hablamos de comunicación bidireccional se refiere a la capacidad de una plataforma para facilitar la comunicación en ambas direcciones entre una infraestructura en la nube y dispositivos IoT (Internet de las Cosas)

Azure Event Hub - Componentes

1. **IoT Hub** El núcleo de la plataforma para la comunicación entre dispositivos IoT y la nube.
2. **Dispositivos** Equipos o sensores IoT que envían datos y reciben comandos.
3. **Device Twins** Registro digital que mantiene las propiedades y el estado deseado de un dispositivo.
4. **Module Twins** Similar a los Device Twins, pero para módulos dentro de un dispositivo IoT.
5. **Mensajes** Datos y comandos enviados entre dispositivos e IoT Hub.

Casos de Uso:
Gestión de dispositivos IoT.
Transmisión de datos de telemetría.
Ejecución de comandos en dispositivos.

¿Qué es Azure Stream Anlytics?

Es un servicio en la nube para el procesamiento y análisis de datos en tiempo real. Permite transformar, filtrar y agregar datos a medida que fluyen desde diversas fuentes, proporcionando insights en tiempo real para la toma de decisiones rápida.

Azure Event Hub - Componentes

1. **Entradas** Fuentes de datos desde las cuales el servicio ingesta datos en tiempo real, como Azure Event Hubs, Azure IoT Hub y Azure Blob Storage.
2. **Consultas SQL-like** Lenguaje similar a SQL utilizado para definir cómo se procesan los datos. Incluye operaciones de filtrado, transformación y agregación.
3. **Salidas** Destinos para los datos procesados, como bases de datos, almacenamiento en la nube y servicios de visualización.
4. **Configuración de Recursos** Ajuste de la capacidad de procesamiento mediante unidades de streaming, que determinan el rendimiento del servicio.
5. **Stream Analytics Job** Una instancia específica para un servicio específico del servicio Azure Stream Analytics. Un trabajo (job) configura y ejecuta el procesamiento de datos en tiempo real, incluyendo entradas, consultas y salidas.

Casos de Uso:
Procesamiento de datos de telemetría.
Detección de patrones y generación de alertas.
Visualización en tiempo real

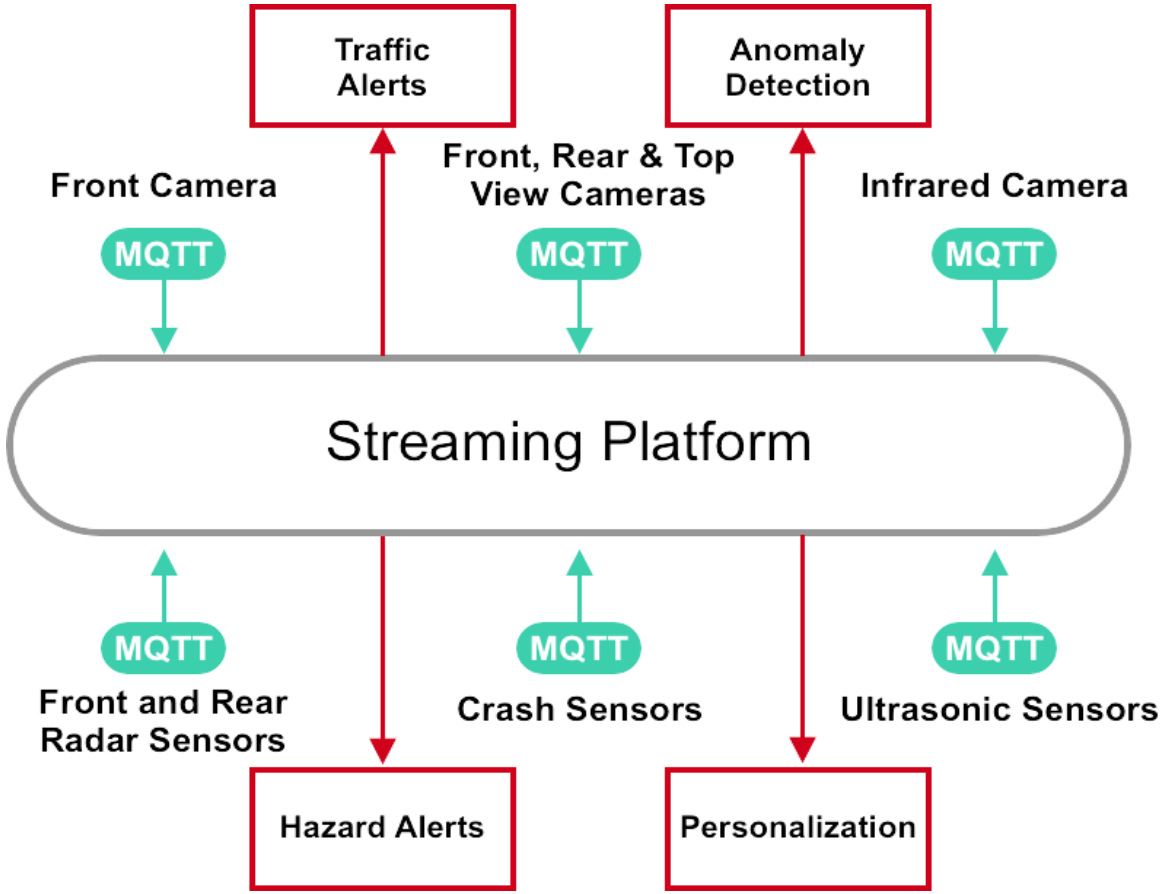
Comparativo de Servicios

Característica	Azure Event Hubs	Azure IoT Hub	Azure Stream Analytics
Descripción General	Servicio de ingesta de eventos en tiempo real	Plataforma para la gestión y comunicación de dispositivos IoT	Servicio de procesamiento y análisis de datos en tiempo real
Propósito Principal	Recibir y procesar grandes volúmenes de eventos y datos	Gestionar la comunicación bidireccional con dispositivos IoT	Procesar y analizar flujos de datos en tiempo real
Tipo de Datos	Eventos y datos de telemetría de aplicaciones	Datos de telemetría y mensajes de dispositivos IoT	Datos en tiempo real de eventos y flujos de datos
Escalabilidad	Alta escalabilidad, maneja millones de eventos por segundo	Soporta millones de dispositivos IoT	Escala según unidades de streaming configuradas para el trabajo
Persistencia de Datos	Almacenamiento temporal; opción de Capture en Blob Storage	No proporciona almacenamiento de datos por sí mismo	No almacena datos; se enfoca en el procesamiento en tiempo real
Seguridad	Autenticación mediante Azure Active Directory y SAS	Autenticación con certificados y gestión de identidades	Seguridad en el acceso a datos y servicios mediante configuraciones y permisos
Integración	Se integra con Azure Stream Analytics, Azure Functions, otros servicios	Se integra con Azure Stream Analytics, Azure Functions, otros servicios	Se integra con Event Hubs, IoT Hub, y otros servicios para procesamiento de datos
Componentes Clave	Event Hubs, Particiones, Consumer Groups	IoT Hub, Dispositivos, Device Twins, Módulos de Dispositivo	Consultas SQL-like, Ventanas de Tiempo, Funciones de Agregación
Casos de Uso Comunes	Análisis de eventos en tiempo real, telemetría de aplicaciones	Gestión de dispositivos IoT, comunicación bidireccional, telemetría	Transformación y análisis de datos en tiempo real, detección de patrones
Costo	Basado en el número de eventos y particiones	Basado en el número de dispositivos y mensajes	Basado en unidades de streaming y la complejidad de las consultas

Detección de fraude en tiempo real

- Actuar en tiempo real
- Detectar fraude
- Minimizar el riesgo
- Mejorar la experiencia del cliente

Automotor



El futuro de la automotriz.

La industria es un clúster de datos en tiempo real.

Cliente 360

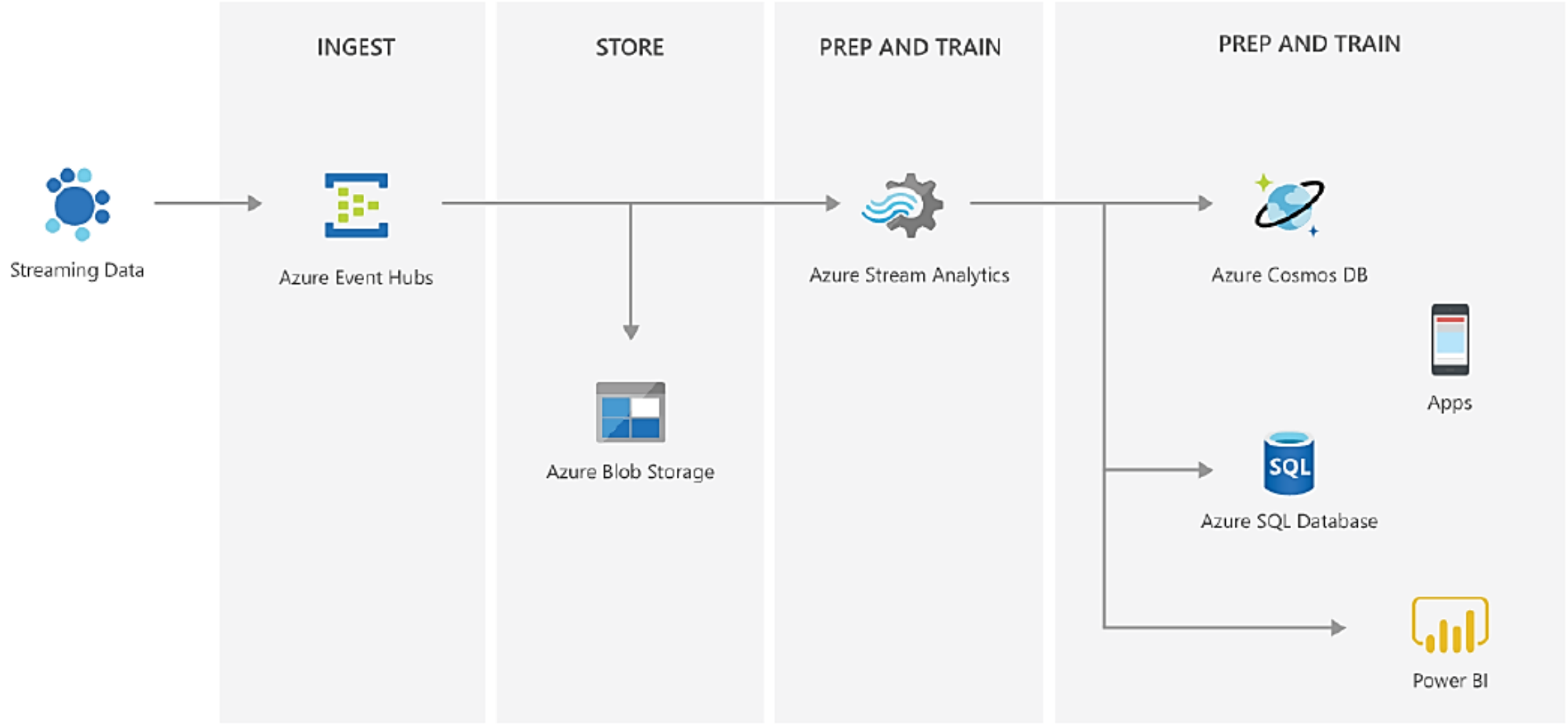
- Integración de datos mejorada
- Incrementar las oportunidades de ventas adicionales y cruzadas
- Incrementar la escalabilidad y flexibilidad
- Ahorro de costos

Comercio electrónico en tiempo real

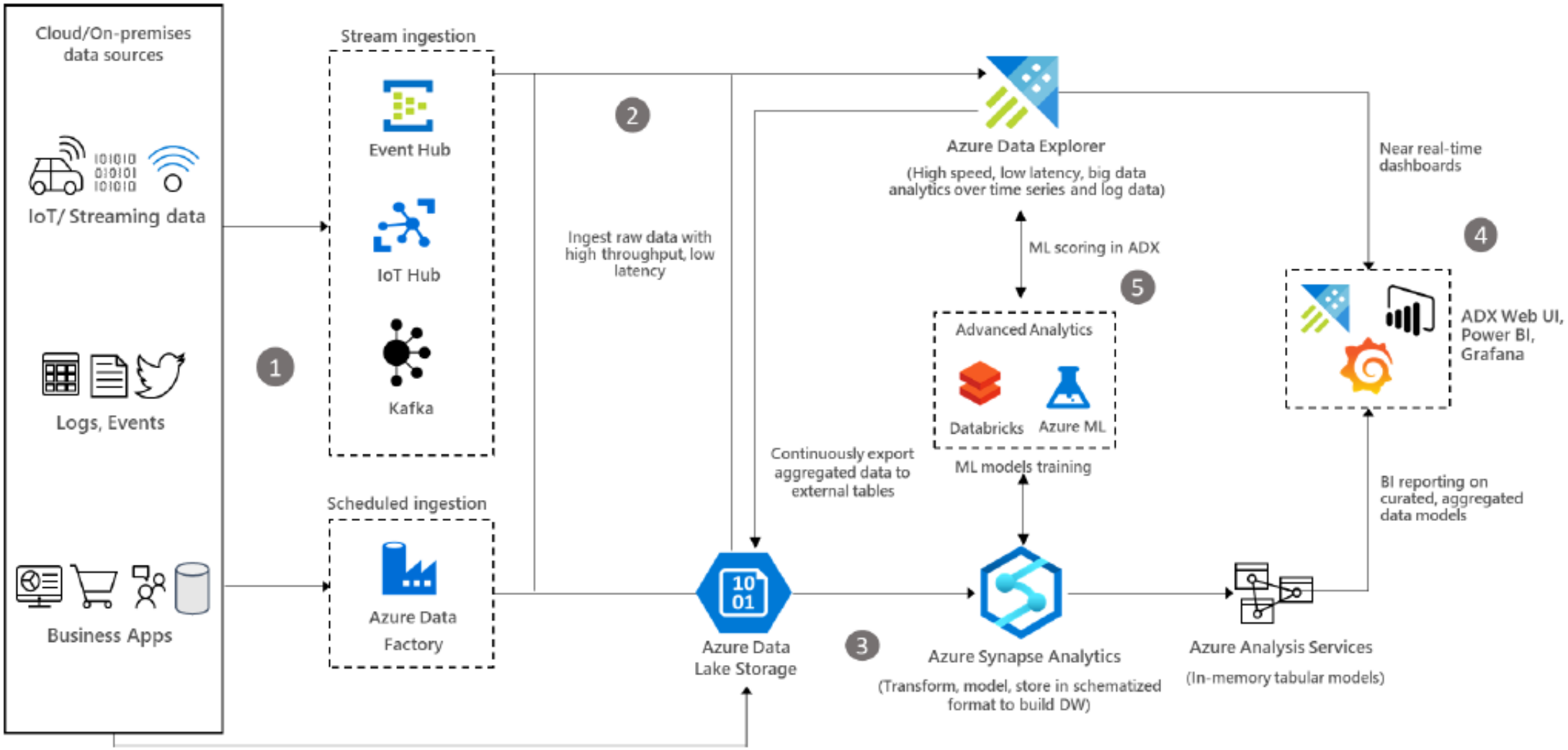
Programa de recompensas

- Incorporación de nuevos comerciantes más rápido
- Mayor velocidad a la que se entregan las aplicaciones móviles a los clientes
- Habilitó una vista completa de 360° de los clientes
- Supervisión y rendimiento mejorados
- Ahorros proyectados de millones de dólares

Streaming sin Servidor con Event Hubs y Stream Analytics



Arquitectura Lambda – Analítica Avanzada

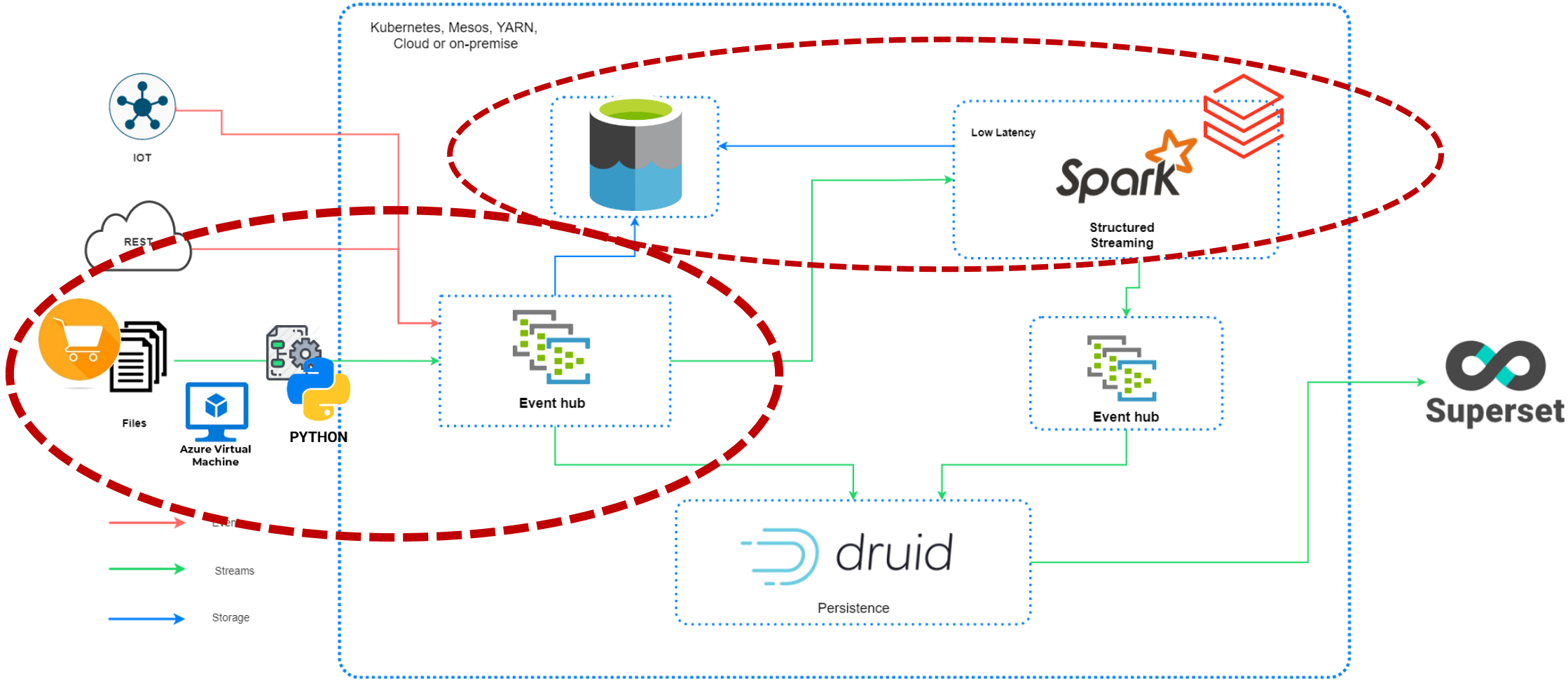


LABORATORIO: Ingesta en Real Time con Databricks

Caso de Uso – Análisis y acción operativa en tiempo real

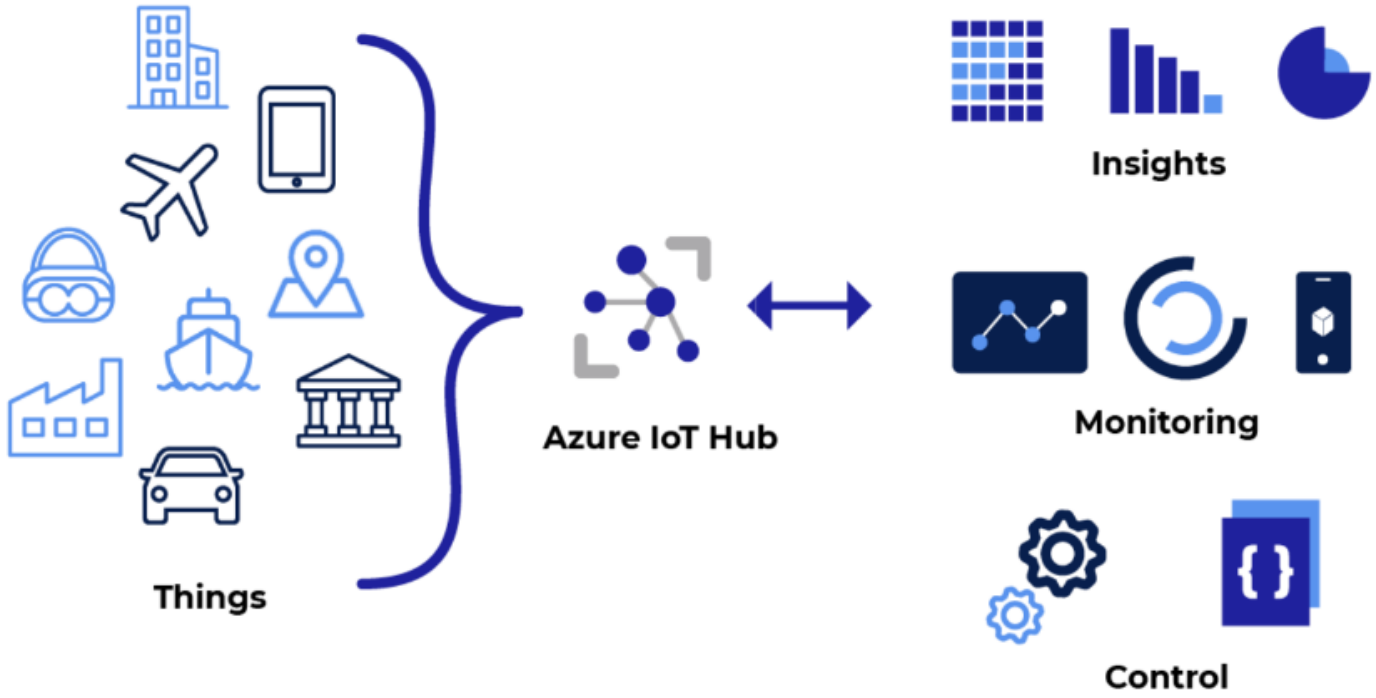
Una analista de negocio de una empresa retail desea analizar los datos de navegación generados por su plataforma ecommerce en tiempo real y realizar una acción operativa.

Real Time Ingest



¿Qué es IoT?

IoT (Internet of Things) es una red de objetos físicos incorporados con sensores, software y otras tecnologías para conectar e intercambiar datos con otros dispositivos y sistemas a través de Internet.



Aplicaciones de IoT

1. **Domestica:** Control y automatización del hogar (iluminación, calefacción).
2. **Salud:** Monitoreo de pacientes a distancia (dispositivos de salud conectados).
3. **Agricultura:** Monitoreo de cultivos y ganado (sensores de humedad, temperatura del suelo).
4. **Industria:** Mantenimiento predictivo, control de procesos industriales.
5. **Ciudades Inteligentes:** Gestión del tráfico, control de iluminación pública.

IOT

Dispositivos Inteligentes

Son objetos físicos equipados con sensores, procesadores y software que les permiten recopilar datos del entorno, tomar decisiones basadas en esos datos y comunicarse con otros dispositivos o sistemas a través de redes.

Características:

1. **Sensores:** Capturan información del entorno (como temperatura, humedad, luz, movimiento).
2. **Actuadores:** Realizan acciones basadas en las decisiones del dispositivo (como encender una luz o ajustar un termostato).
3. **Procesamiento Local**: Capacidad de procesar datos localmente y tomar decisiones en tiempo real sin necesidad de enviar toda la información a la nube.
4. **Conectividad**: Permiten la comunicación con otros dispositivos y sistemas a través de redes (Wi-Fi, Bluetooth, Zigbee, etc.).
5. **Autonomía:** Funcionan de manera autónoma para realizar tareas específicas basadas en algoritmos y datos recogidos.

IOT

Conectividad en IoT

Se refiere a la capacidad de los dispositivos inteligentes para intercambiar datos y comunicarse entre sí y con sistemas externos a través de diferentes tipos de redes.

Características:

1. **Protocolos de Comunicación:**

   a) HTTP/HTTPS: Protocolos basados en la web utilizados para la comunicación entre dispositivos y servidores.
   b) MQTT (Message Queuing Telemetry Transport): Protocolo ligero para la transmisión de mensajes entre dispositivos en redes de baja ancho de banda.
   c) CoAP (Constrained Application Protocol): Protocolo diseñado para dispositivos con recursos limitados.

2. **Redes y tecnologías:**

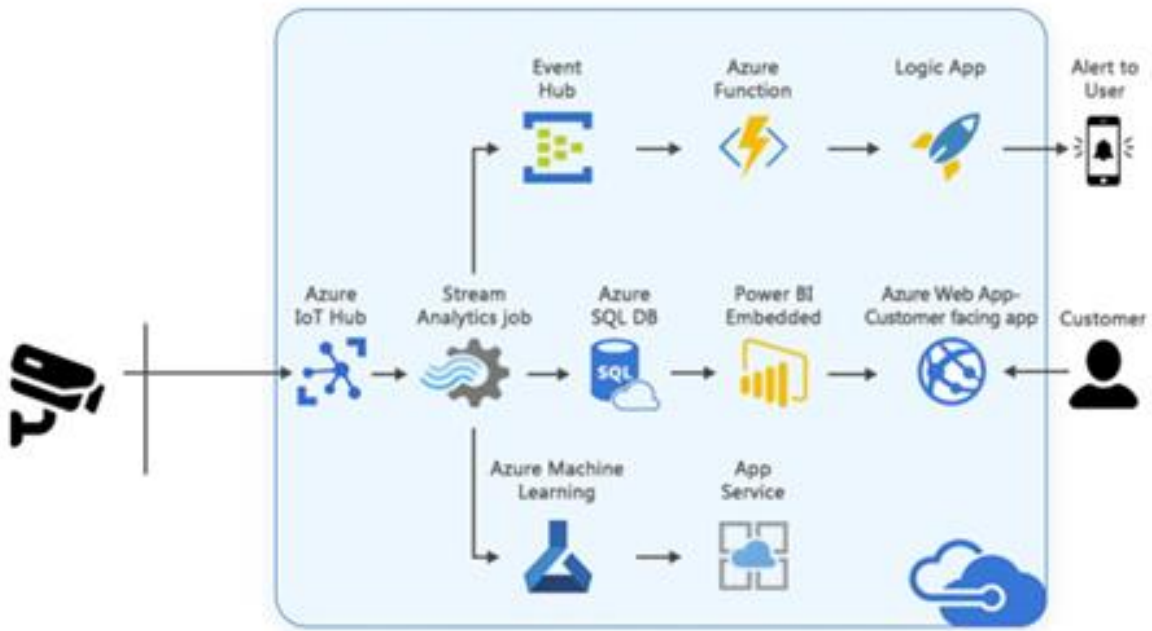
   a) Wi-Fi: Conexión inalámbrica común en dispositivos domésticos.
   b) Bluetooth y Bluetooth Low Energy (BLE): Utilizado para comunicación de corto alcance y bajo consumo de energía.
   c) Zigbee y Z-Wave: Protocolos diseñados para redes de área personal y aplicaciones de automatización del hogar.
   d) Redes Móviles (4G/5G): Para la conectividad de dispositivos en movimiento o en áreas sin cobertura Wi-Fi.

IOT

3. Cloud Computing:

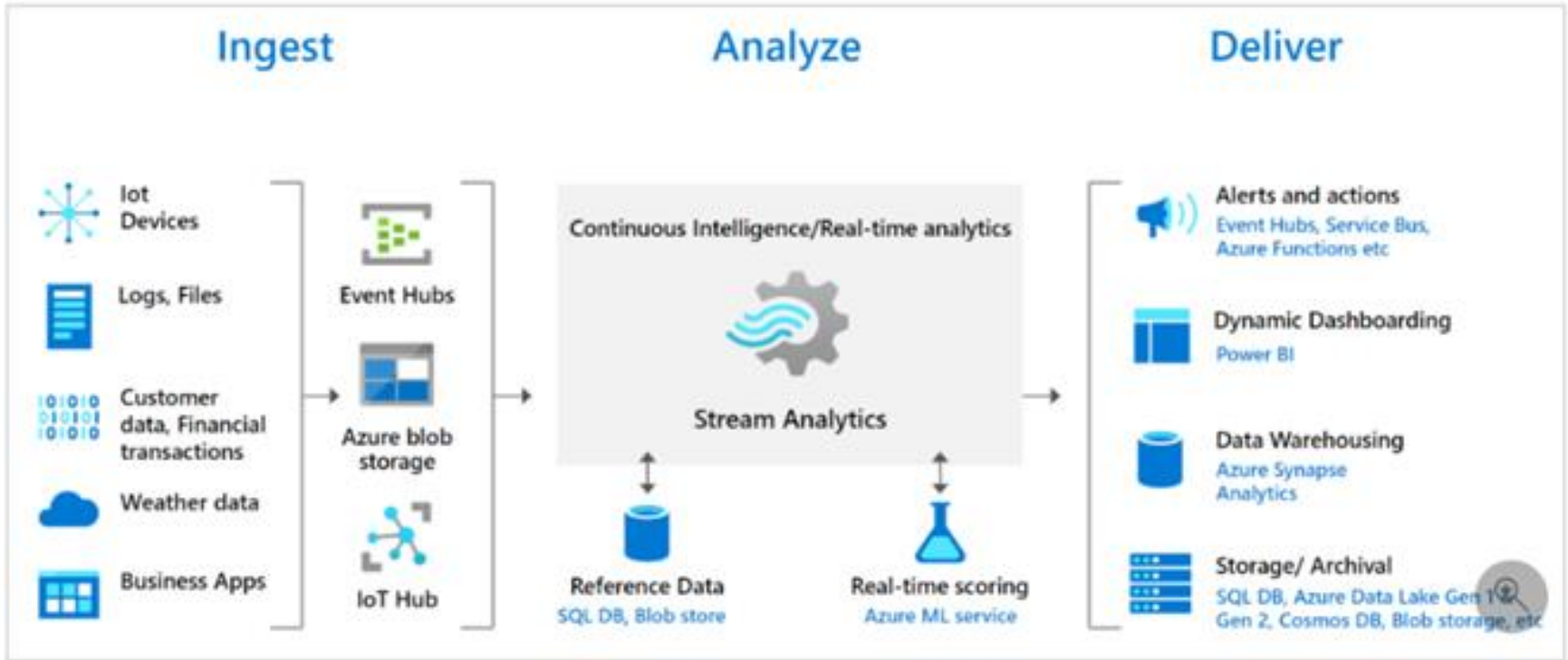
a) Servicios en la Nube: Permiten la integración de dispositivos IoT con plataformas en la nube para el almacenamiento de datos, procesamiento y análisis.

b) Edge Computing: Procesamiento de datos cerca del lugar donde se generan para reducir la latencia y el ancho de banda necesario.



Arquitectura de Ejemplo: Sistema de Monitoreo en Tiempo Real de Sensores IoT

Descripción del Caso de Uso:
Imagina un sistema de monitoreo en tiempo real para una red de sensores IoT instalados en una planta industrial. El objetivo es captar datos de sensores en tiempo real, procesarlos para detectar anomalías, y visualizar los resultados en un panel de control.



Arquitectura de Ejemplo: Sistema de Monitoreo en Tiempo Real de Sensores IoT

IoT Hub: Gestión y Recepción de Datos de Sensores

Azure IoT Hub actúa como el punto de entrada principal para todos los datos generados por los sensores IoT. Cada sensor envía sus datos de telemetría a IoT Hub, donde se **recibe** y **gestiona** de manera segura.

Operación: Los dispositivos IoT se conectan a IoT Hub mediante protocolos estándar como MQTT o HTTP. IoT Hub asegura la autenticación y la comunicación segura entre los dispositivos y la nube.

Event Hub: Ingesta de Datos para Procesamiento en Tiempo Real

Event Hub recibe los datos de telemetría desde IoT Hub. Event Hub actúa como un buffer que permite la ingesta y el procesamiento de grandes volúmenes de datos en tiempo real.

Operación: IoT Hub enruta los datos de los dispositivos a Event Hub a través de rutas predefinidas. Event Hub organiza los datos en particiones para permitir la escalabilidad y el procesamiento paralelo.

Arquitectura de Ejemplo: Sistema de Monitoreo en Tiempo Real de Sensores IoT

Azure Stream Analytics Job: Procesamiento en Tiempo Real

Un Azure Stream Analytics Job procesa los datos en tiempo real que llegan a Event Hub. Utiliza consultas SQL-like para analizar y transformar los datos, identificar patrones, y detectar anomalías.

Operación: La consulta definida en el job de Stream Analytics puede filtrar, agrupar, o realizar cálculos sobre los datos de sensores. Por ejemplo, puede calcular promedios de temperatura y detectar si un sensor reporta una temperatura anormalmente alta.

Azure Stream Analytics: Salidas y Visualización de Resultados

Los resultados procesados por Azure Stream Analytics se envían a destinos como Azure Blob Storage para almacenamiento a largo plazo, a Power BI para visualización en tiempo real, o a bases de datos para análisis adicionales.

Operación: Dependiendo de la configuración de salida, los resultados de las consultas pueden ser enviados a un dashboard en Power BI para que los operadores puedan ver datos en tiempo real, o almacenados en Azure Blob Storage para análisis histórico.

Resumen de la Arquitectura

IoT Hub

Rol Recepción y gestión segura de datos de sensores IoT.
Datos Telemetría de sensores, como temperatura, humedad, etc.
Función Enviar datos a Event Hub.

Event Hub

Rol Ingesta y almacenamiento temporal de grandes volúmenes de datos.
Datos Datos de telemetría provenientes de IoT Hub.
Función Enviar datos a Azure Stream Analytics para procesamiento.

Azure Stream Analytics Job

Rol Procesamiento y análisis de datos en tiempo real.
Datos Datos de telemetría de Event Hub.
Función Ejecutar consultas en tiempo real para detectar anomalías y patrones.

Azure Stream Analytics

Rol Envío de resultados procesados a destinos finales.
Datos Resultados de las consultas, como alertas o datos analizados.
Función Visualización en Power BI, almacenamiento en Azure Blob Storage, o envío a bases de datos para análisis adicional.

Stream Analytics

En Microsoft Azure, **Stream Analytics** es un servicio de análisis en tiempo real que permite procesar y analizar datos a medida que se generan. Dentro del contexto de Stream Analytics, el término "Stream Analytics Job" se refiere a la **configuración de trabajos específicos que puedes crear para procesar datos en tiempo real**. Aunque el servicio en sí se centra en la creación de trabajos de análisis de flujo, los "servicios" específicos de Stream Analytics están relacionados con los componentes y características que puedes usar dentro de estos trabajos.

Azure Event Hubs: Servicio de ingesta de eventos que permite recibir grandes volúmenes de datos en tiempo real.

Azure IoT Hub: Plataforma para la comunicación bidireccional entre dispositivos IoT y la nube.

RONDAS DE PREGUNTAS

¡GRACIAS!