Propuesta Técnica
Plataforma de Analítica Avanzada -Belcorp

Autor: Beto Castillo, Data & A.I Architect – Indra Perú

Version 1.0

1. Introducción

El documento técnico presenta una arquitectura de analítica avanzada diseñada para integrarse con la plataforma de Belcorp. A diferencia de enfoques que proponen soluciones aisladas, nuestro objetivo es construir una arquitectura que extienda y aproveche al máximo los activos actuales de datos, asegurando interoperabilidad, gobernanza y escalabilidad.

2. Enfoque Tecnológico

Es importante aclarar que **la propuesta presentada no está alineada a un stack tecnológico específico**, ya que actualmente no contamos con el detalle completo de la infraestructura de datos y servicios de Belcorp. Sin embargo, lo descrito más delante es **intencional y estratégico.**

En lugar de asumir tecnologías concretas, hemos optado por presentar una arquitectura basada en componentes funcionales que componen cualquier plataforma moderna de analítica avanzada.

Estos componentes han sido diseñados siguiendo las mejores prácticas de la industria y son completamente agnósticos, lo que permite adaptarlos de forma flexible a distintos entornos; ya sea en nube pública, híbrida o infraestructura on-premise, así como a distintos proveedores como AWS, Azure, GCP o tecnologías open source.

El enfoque de esta propuesta no es imponer una solución tecnológica cerrada, sino definir los pilares fundamentales que debe tener una plataforma robusta, escalable y gobernada para el desarrollo, despliegue y monitoreo de modelos de machine learning en producción.

Una vez tengamos visibilidad del stack actual de Belcorp, nuestra propuesta puede ajustarse y optimizarse para:

- Integrarse eficientemente con la arquitectura de datos existente.
- Reutilizar activos tecnológicos disponibles.
- Respetar lineamientos de seguridad, gobernanza y políticas de TI.
- Minimizar el time-to-value y la complejidad operativa.

3. Objetivo

Construir una plataforma de analítica avanzada alineada a la arquitectura de datos de Belcorp, que permita:

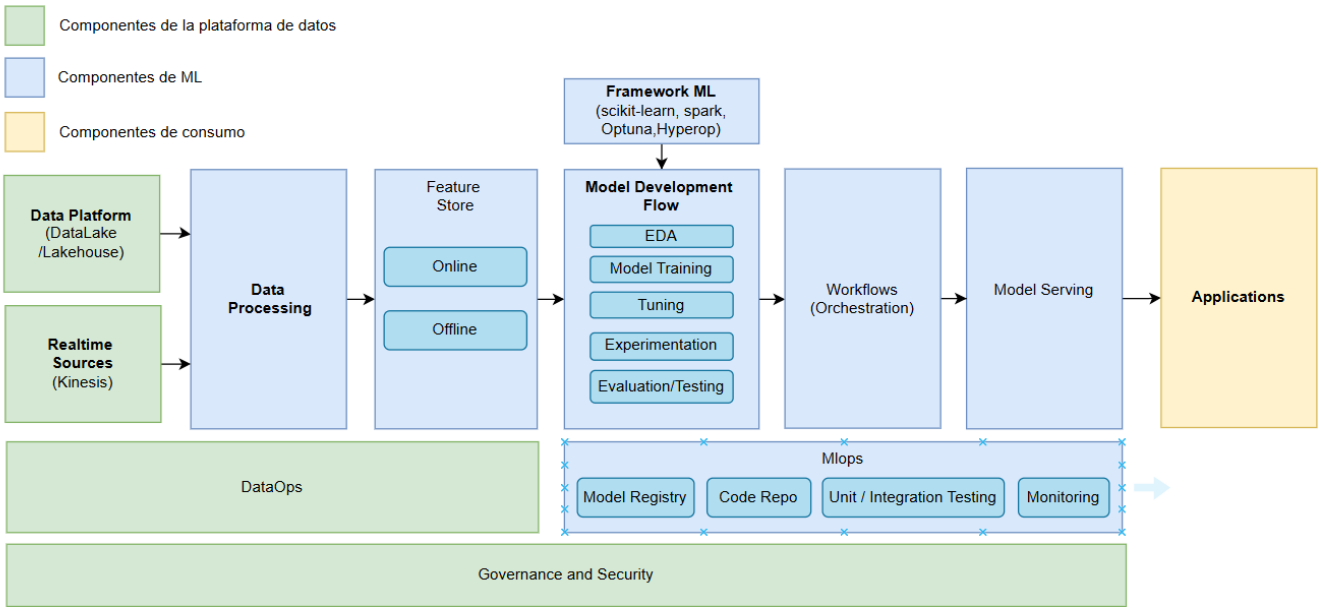
- Agilidad en el desarrollo y despliegue de modelos.
- Control y trazabilidad total de los ciclos de vida.
- Seguridad y gobernanza transversal.
- Integración con las aplicaciones del negocio.

4. Beneficios Esperados

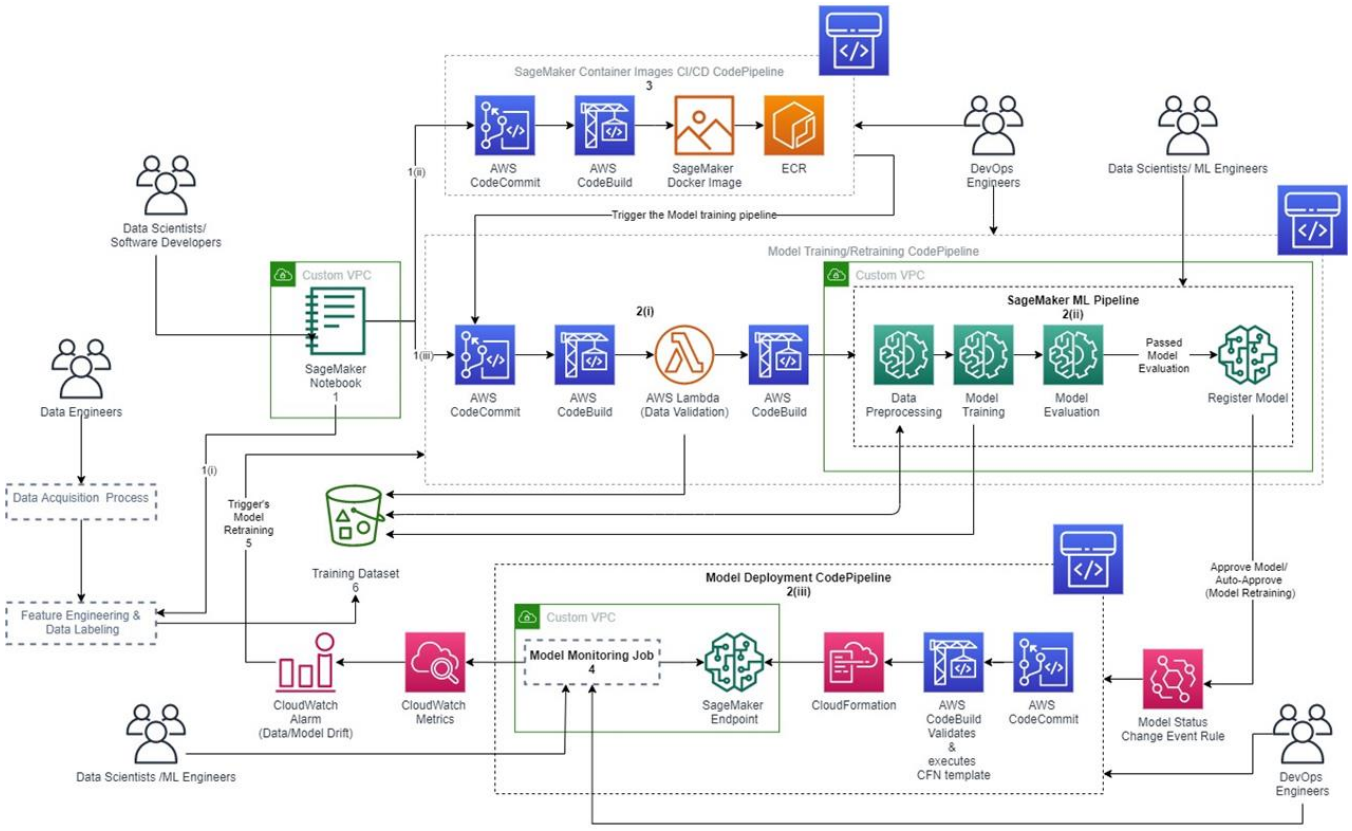
Beneficio	**¿Cómo se logra?**
Aceleración del time-to-value	Aprovechando la plataforma de datos existente
Simplicidad y bajo acoplamiento	Uso de componentes modulares y reutilizables
Alta gobernanza y control	Control de accesos, linaje,auditoria,etc.
Sostenibilidad a largo plazo	Arquitectura extensible, automatizable y auditable

5. Arquitectura

5.1. Componentes



5.2 Física (Referencial)



6. Descripción de Componentes Propuestos

La arquitectura propuesta se compone de un conjunto de bloques funcionales diseñados para cubrir de forma integral el ciclo de vida de modelos de machine learning en producción. Cada componente cumple un rol específico dentro de la plataforma y busca integrarse con los sistemas existentes de datos y operación.

- **Data Platform (DataLake / Lakehouse):** Almacén centralizado para datos estructurados y no estructurados, utilizado como fuente principal para el entrenamiento y monitoreo de modelos.
- **Realtime Sources:** Fuente de datos en tiempo real utilizada para alimentar flujos de inferencia y modelos de respuesta rápida.
- **Data Processing:** Responsable de la limpieza, validación y transformación de datos para preparar insumos de calidad para la analítica avanzada.
- **Feature Store (Online / Offline):** Repositorio centralizado de variables predictoras reutilizables, que garantiza coherencia entre entrenamiento e inferencia.
- **Model Development Flow:** Flujo que incluye análisis exploratorio (EDA), entrenamiento, búsqueda de hyperparametros , tracking y evaluación de modelos.
- **Framework ML:** Ecosistema de herramientas para abstraer, estandarizar y promover la reutilización de código, tomando como base librerías como scikit-learn, Spark MLlib, Optuna y HyperOp.
- **Workflows (Orchestration):** Orquestación automatizada de pipelines ML, permitiendo ejecución secuencial o paralela, manejo de errores y programación periódica.
- **Model Serving:** Infraestructura que permite exponer modelos entrenados como servicios, accesibles por otras aplicaciones mediante APIs o procesos batch.
- **Applications:** Aplicaciones del negocio que consumen las predicciones de los modelos para mejorar decisiones y procesos operativos.
- **Model Registry:** Sistema de versionamiento y gestión de modelos con trazabilidad y estados de publicación (Staging, production, archived)
- **Code Repo:** Repositorio de código fuente donde se gestionan scripts, notebooks y pipelines de datos y modelos.
- **Unit / Integration Testing:** Pruebas automatizadas que aseguran la calidad del código y la funcionalidad de los pipelines antes del despliegue.
- **Monitoring:** Sistema de observabilidad del rendimiento del modelo en producción, incluyendo métricas, alertas y detección de drift.
- **Governance and Security:** Capa transversal que garantiza cumplimiento de normativas, control de accesos, auditoría y protección de datos.
- **DataOps & MLOps:** Buenas prácticas y herramientas para operacionalizar datos y modelos de forma continua, confiable y automatizada.

7. Conclusión

Esta arquitectura no solo propone una solución técnica robusta, sino que ofrece una visión estratégica y sostenible de analítica avanzada. El resultado será una plataforma que crece con el negocio, sin reinventar lo ya construido, y que permite llevar modelos de ML a producción de forma rápida, controlada y escalable.