**Lectura1:**

**Migración de almacenamiento de Messenger para optimizar el rendimiento**

Cuando se diseñó originalmente Messenger era muy similar a un correo electrónico, en estos tiempos es un sistema de comunicaciones en real time, móvil y novedoso. Para llegar a tener estos cambios por ejemplo el servicio monolítico se separó en almacenamiento cache de lectura completa para consultas; y Iris para poner en cola las escrituras de los suscriptores y un servicio de almacenamiento.

Moderanizando el servicio de almacenamiento; para que sea mas rápido, mas confiable, mas eficiente. Cambios:

\*Rediseño y simplificación de esquema de datos.

\*Se pasó de HBase a MyRock

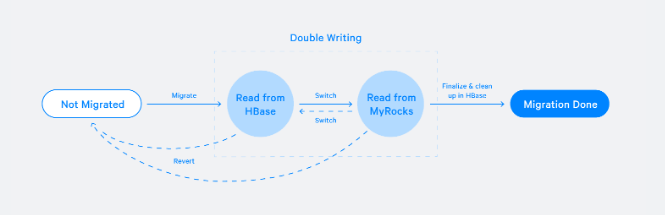
\*De almacenar en BD en discos giratorios a flashear en SKU de Lightning Server

También se ha mejorado la resistencia del sistema, la latencia se redujo y el consumo de almacenamiento se redujo al 90%.

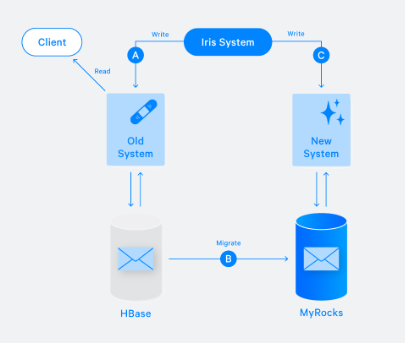
**Manejando el reto de migrar a escala:**

Pasar de HBase a MyRocks dio varios beneficios. Aprovechando el SKU de Lightning Server del proyecto Open Compute se pudo proporciona datos alojados en almacenamiento flash en lugar de depender de discos giratorios. El traspaso no fue nada fácil o un verdadero reto porque no se podría afectar a la experiencia del cliente que usaba dicha aplicación.

Para enfrentar los desafíos, se diseño dos flujos de migración para todas las cuentas de Messenger. El flujo normal cubrió 99.9% de las cuentas, mientras que el flujo de migración en bufer cubrió las cuentas restantes difíciles de migrar. Cuando se estuvo seguro de la migración se desconecto el sistema anterior.

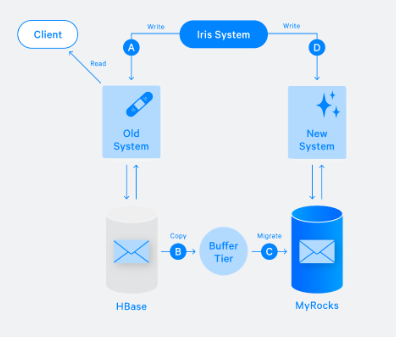


**Migración Normal**

Registrar el ultimo dato en los dos sistemas nuevo y anterior para la validación de los datos se confirma si los datos de HBase y MyRocks coinciden. Y se valida el API, emitiendo consultas a ambos sistemas. Antes de finalizar se verifica si la cuenta migro correctamente. 

Flujo de Migración de almacenado

Algunas cuentas no se podian migrar de forma normal, en general las cuentas de empresas tenian estos problemas. Se establecio una hora limite (inicio) para poder iniciar la migración de los datos.



**Migrar a escala.**

Se aprovechó el marco de Bistro, cuando se migro las cuentas al nuevo sistema. Todas las cuentas creadas nuevas irán al nuevo sistema, gradualmente se deja de utilizar HBase, un cluster a la vez.

**Beneficios del Nuevo Sistema**

99.9% se migro en dos semanas en la migración normal y el resto del flujo en migración amortiguada. MyRocks esta optimizando tanto en lectura como en escritura, al aprovechar Lighttning con almacenamiento flash, se gana latencia siendo 50 veces mas baja que el sistema anterior.

Estas mejoras en rendimiento y la arquitectura han permitido que sea mas flexible agregar nuevas características. Nos permite buscar mensajes de contenido móvil con mas rapidez.

-------

<https://code.fb.com/core-data/migrating-messenger-storage-to-optimize-performance/>

**Lectura 2:**

**Introducing The Million Playlist Dataset and RecSys Challenge 2018**

La Digital Music Alliance en su informe anual 2018, dicen que el 54% consumidores dicen que las play list están reemplazando a los álbumes. Pero a muchos usuarios no tan solo les gusta escuchar las play list sino también crearlas y categorizarlas según diferentes aspectos.

Otras de las cosas que gustan a la gente es la búsqueda de lista de reproducción, al aprender con las listas que otras personas crean y ver la relación de la persona con la música y que emojis utilizan las personas para describir una lista de reproducción.

Al descubrir más la naturaleza de las listas de reproducción, también se puede sugerir otras pistas a las personas y ayudar a encontrar la música que aman.

Para que se de este tipo de investigación a escala, se lanzó el The Million Playlist Dataset (MPD) a la comunidad de investigación academica. Un conjunto de datos de 1 millon de play list que consta de más de 2 millones de canciones únicas de casi 300,000 artistas. Representando el conjunto de datos de play list más grande del mundo. Las listas fueron creadas por los usuarios entre 2010 y noviembre 2017.

Se asociaron con investigadores de la Universidad Johannes-Kepler de Linz y la Universidad de Massachusetts Amherst para lanzar el desafio RecSys 2018, que es el desafio anual de la ciencia de datos para la conferencia de sistemas de recomendaciones de ACM. La tarea es un verdadero reto donde se le pide a los investigadores que envíen sus sistemas para predecir correctamente las canciones que se han ocultado de la lista de reproducción reales creadas por los usuarios.

<https://labs.spotify.com/2018/05/30/introducing-the-million-playlist-dataset-and-recsys-challenge-2018/>

**Lectura 3:**

**Scaling Machine Learning at Uber with Michelangelo**

En este articulo se reflexiona de la evolución que tuvo el machine learning en Uber

Cero a 100 en tres años:

Hace tres años no se usaba mucho el machine learning en Uber pero a medida que la empresa crecía y los servicios se hacían más complejos surgió la oportunidad de utilizar ML. Donde con el paso del tiempo se fueron sofisticando hasta llegar a tener herramientas e infraestructuras de ML avanzadas y muchos componentes de intregración.

Casos de uso de ML en Uber:

Uber come: Modelos de Ml sugieren restaurantes y elementos de Menú basados en datos históricos e información de la sesión actual del usuario en la app (ejm: su consulta de búsqueda).

Pronostico del Mercado: Modelos espaciotemporales de pronostico que son capaces de predecir donde estará la demanda de usuarios y la disponibilidad de conductores y socios en varios lugares y momentos en el futuro. Uber puede alentar a los socios conductores a que vayan a un lugar donde habrá demanda futura.

Atención al Cliente: Los modelos ML están ayudando a solucionar problemas para dar respuestas automáticas a las preguntas, quejas,etc que hacen los clientes.

Cheque de Viaje: asistencia por un choque al reconocer una demora en el viaje ya que se tiene acceso a la información del GPS.

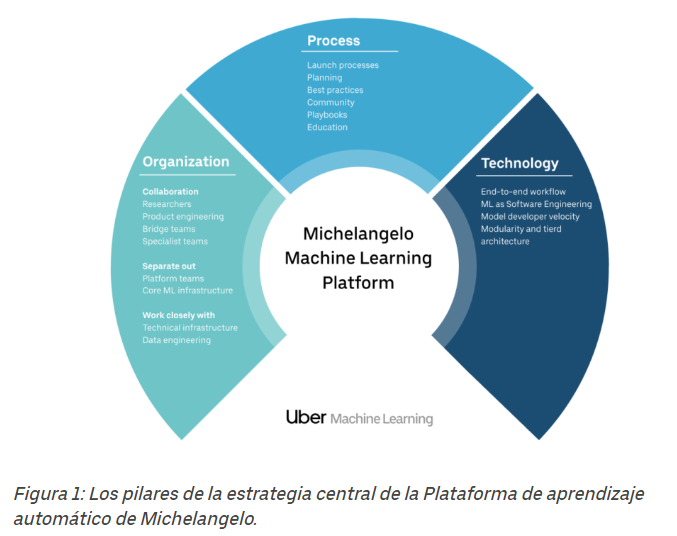
Tiempos estimados de llegada ETA: es muy difícil de predecir, las ETA precisas son fundamentales para que el usuario tenga una experiencia positiva. Se utilizó un modelo de ML para poder predecir los errores de calculo y así nuevamente el ETA corregido.

Chat con un clic: se utiliza el modelo de procesamiento de lenguaje natural NLP

Coches de auto-conducción: ML para la identificación de objetos y la planificación de movimientos.

**Como escalamos ML en Uber**

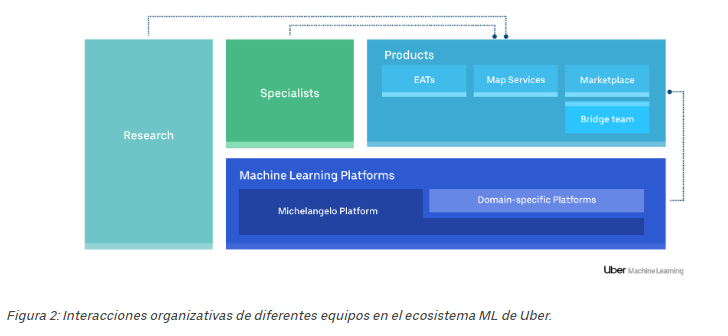
Democratizando las herramientas y aistencia del equipo técnico, optimizando la velocidad del desarrollador, la ingeniería de software y la flexibilidad del sistema.

La escala exitosa de ML también existen consideraciones organizacionales y diseño de proceso. Factores críticos explicados en tres pilares: organización, proceso y tecnología. 

**Organización**

Lograr que las personas adecuadas trabajen en los problemas correctos ha sido fundamental para crear soluciones de alta calidad y desplegarlas de manera consistente y exitosa en la producción.

Uber actualmente opera con los siguientes roles y responsabilidades principales:



A medida que el sistema se vuelve más sofisticado y los problemas que se resuelven son más complejos, la demanda crece por flexibilidad adicional, extensibilidad y experiencias de desarrollo de ML específicas de dominio.

La escala de ML de alta calidad en toda la empresa requiere una organización conectada y colaborativa. La educación permanente es importante para el equipo de ML y deben estar en la vanguardia de las nuevas investigaciones. Se organzición sesiones de capacitación especial con los empleados que usan estas tecnologías.

Tecnologia:

Flujo de trabajo de extremo a extremo: se necesita soporte para todo el flujo de trabajo desde la administración de BD, construir modelos, desarrollo de ML, implementación de modelos, realizar predicciones y supervisión de prediciones.

ML como ingeniería de software: es valioso trazar analogías entre el desarrollo de ML y el desarrollo de software, luego aplicar patrones de herramientas y metodologías de desarrollo de software al modelo ML

Velocidad del desarrollador de modelos: el desarrollo de ML es un proceso muy iterativo: la innovación y los modelos de alta calidad provienen de muchos y muchos experimentos. Debido a esto, la velocidad del desarrollador de modelos es muy importante

Modularidad y arquitectura escalonada: proporcinar flujos de trabajo de extremo a extremo es muy importante para manejar casos de uso de ML más comunes, pero para abordar los casos menos comunes y más especializados, es importante contar con componentes primitivos que se puedan ensamblar de manera específica.

**Conferencia 1**

En todas las organizaciones se esta tratando de encontrar y estudiar el gran volumen de la data, y hay personas que se encuentran buscando especializarse en ella. Y Kaiser esta experimentando con AI para personalizar la intervención asistencial.

Las investigaciones en Boston Children Hospital lograron aplicarar algoritmos para predecir el Ebola una semana antes que suceda. Analizando diversas fuentes de datos como imágenes, exámenes, etc. con la esperanza de lograr descubrimientos de drogas mas rápido para identificar las mutaciones que causan particulares enfermedades que reducen en la declaración medica para predecir y tratar la enfermedad antes de su manifestación para mejorar los resultados del paciente y probar un tratamiento personalizado.

Esta revolución es como una gran explosió in la Genomics Data. En 2025 la gran cantidad de Genomics Data podria predecir en menor tiempo el diagostico.

El costo de procesamiento de la secuencia del ADN en una año es creciendo exponencialmente año tras año que son en base a procesamiento no escalables, los datos para mejorar la vida de las personas son frágiles y primitivos. En la actualidad se han tomado algoritmos para una tubería de 40 etapas que luego servirán para hacer análisis y técnicas de performans. Se crearan profesiones para la analizar la data según las especialidades medicas.

En la actualidad existen soluciones rapidas de tuberías, técnicas analíticas, inteligencia artificial y colaboradores especializados. Para construir en Spark y optimizarlo. Descubriendo nuevas drogas para el tratamiento y ayudar a que las personas se sanen.

El machine learning aprende los recién nacidos de sus parentesco para poder predecir si podrían encontrarse con una enfermedad que se pueda heredar, estas variables pertenecen a la historia clínica. Procesos ETL para cargar la data e integrarla al ecosistema de big data y obtener la data

**Conferencia 2**

**BUILDING THE FUTURE OF DRUG DISCOVERY**

El doctor Jeff Reid, del centro de genética de Regeneron. Nos cuenta que toda el mundo esta tocado por al menos una enfermedad. Y que en los últimos 30 años, en Regeneron se ha hecho todo lo posible para llevar la ciencia a la medicina para desarrollar terapias que ayuden a mejorar la salud de las personas.

En el 2014 se lanzo el “ Centro de Genetica Regeneron” y se embarcaron en uno de los mayores esfuerzos de secuenciación en el mundo para llevar la revolución genómica al desarrollo de medicamentos basados en la genética humana y validados utilizando las mejores tecnológicas para ratones.

Debido a esa gran escala que se estaban embarcando en el proyecto nació la necesidad de la ciencia de datos para construir y validar el impacto de variantes genéticas. La primera estrategia de datos a gran escala ya ha dado sus frutos con un nuevo objetivo de medicamentos para la enfermedad hepática crónica. Pero sabiendo que el conjunto de datos es muy potente en escala TB en rápido crecimiento de mas de 80B puntos de datos a la plataforma de análisis unificados de Databricks para brindar escalabilidad de modo que se pueda trabajar en todas las áreas de enfermedades simultáneamente y permitiendo que todos los usuarios en los equipos de desarrollo de medicamentos obtengan información del masico volumen de datos diversos.

Tener los datos es el primer paso, lo que permite a los equipos de desarrollo de medicamentos responder preguntas con los datos es como estamos construyendo el futuro del descubrimiento de medicamentos.

-----

https://databricks.com/sparkaisummit/north-america/2018-spark-summit-ai-keynotes-2

**Nuevas lecturas:**

**La arquitectura y la historia de Git: un sistema de control de versiones distribuido**