Introducción a Databricks y el ecosistema Spark

Un viaje a través del procesamiento de datos y Big Data

databricks

IMMUNE
TECHNOLOGY INSTITUTE

Introducción al Procesamiento de Datos y Big Data

En este viaje, exploraremos cómo la explosión de datos ha transformado la manera en que las organizaciones toman decisiones.

Desde los primeros días del procesamiento manual de datos hasta la actualidad, donde manejamos volúmenes de información inimaginables hace unas décadas.

Este cambio no solo ha requerido nuevas tecnologías, sino también un nuevo enfoque en cómo almacenamos, procesamos y extraemos valor de estos vastos océanos de datos.

databricks

IMMUNE
TECHNOLOGY INSTITUTE

Orígenes de MapReduce

La revolución del Big Data comenzó con una idea simple pero poderosa.

En 2004, Jeff Dean y Sanjay Ghemawat de Google publicaron un *paper* sobre MapReduce, un modelo de programación que permitía procesar grandes conjuntos de datos distribuidos de manera eficiente.

Esta innovación fue el catalizador para el desarrollo de tecnologías que podrían manejar la escala y complejidad del Big Data.

[The Friendship That Made Google Huge | The New Yorker]

databricks

IMMUNE TECHNOLOGY INSTITUTE

Nacimiento de Apache Hadoop

Inspirados por MapReduce, Doug Cutting y Mike Cafarella crearon Apache Hadoop en 2006.

Hadoop se convirtió en sinónimo de Big Data, ofreciendo un framework que permitía el almacenamiento y procesamiento distribuido de grandes conjuntos de datos a través de clusters de computadoras y usando modelos de programación simples.

Hadoop democratizó el acceso al procesamiento de Big Data, sentando las bases para la próxima ola de innovaciones que vendrían con posterioridad.

Sin embargo, a pesar de su éxito, Hadoop enfrentó desafíos relacionados con la velocidad de procesamiento y la complejidad en la gestión de recursos. Estas limitaciones fueron particularmente evidentes en aplicaciones que requerían análisis en tiempo real o que involucraban múltiples iteraciones sobre los mismos datos.

databricks

IMMUNE
TECHNOLOGY INSTITUTE

Nacimiento de Apache Hadoop

Inspirados por MapReduce, Doug Cutting y Mike Cafarella crearon Apache Hadoop en 2006.

Hadoop se convirtió en sinónimo de Big Data, ofreciendo un framework que permitía el almacenamiento y procesamiento distribuido de grandes conjuntos de datos a través de clusters de computadoras y usando modelos de programación simples.

Hadoop democratizó el acceso al procesamiento de Big Data, sentando las bases para la próxima ola de innovaciones que vendrían con posterioridad.

Sin embargo, a pesar de su éxito, Hadoop enfrentó desafíos relacionados con la velocidad de procesamiento y la complejidad en la gestión de recursos. Estas limitaciones fueron particularmente evidentes en aplicaciones que requerían análisis en tiempo real o que involucraban múltiples iteraciones sobre los mismos datos.

databricks

IMMUNE
TECHNOLOGY INSTITUTE

Evolución hacia Apache Spark

En 2009, la Universidad de California, Berkeley, presentó Spark, una evolución de la idea original de Hadoop, diseñada para superar sus limitaciones de velocidad.

Spark introdujo el procesamiento en memoria, permitiendo operaciones de datos mucho más rápidas, especialmente para aplicaciones que requieren múltiples iteraciones sobre el mismo conjunto de datos, como los algoritmos de machine learning.

En 2013, el proyecto fue donado a la Apache Software Foundation y se modificó su licencia a Apache 2.0. En febrero de 2014, Spark se convirtió en un **Top-Level Apache Project**

databricks

IMMUNE TECHNOLOGY INSTITUTE

Beneficios de Apache Spark

- Apache Spark revolucionó el análisis de Big Data con tres ventajas clave sobre Hadoop y MapReduce:
  - **Velocidad**, debido a su procesamiento en memoria.
  - **Facilidad de uso**, gracias a sus APIs de alto nivel en Java, Scala, Python y R.
  - **Versatilidad**, al soportar tareas de batch processing, streaming, machine learning y procesamiento de grafos (GraphX), de manera unificada. Estas características lo convierten en una herramienta indispensable para el análisis de datos moderno, eliminando la necesidad de integrar múltiples tecnologías para diferentes tipos de tareas de procesamiento de datos

VS

databricks

IMMUNE
TECHNOLOGY INSTITUTE

Apache Spark en el Mundo del Big Data

La popularidad de Apache Spark ha crecido exponencialmente en la industria.

Empresas de todos los tamaños lo adoptan para resolver problemas complejos de análisis de datos, desde la optimización de cadenas de suministro hasta el desarrollo de modelos predictivos en la salud.

Su capacidad para procesar grandes volúmenes de datos de manera eficiente ha sido un cambio de juego en el mundo del Big Data.

El compromiso continuo con la innovación y mejora, impulsado por una comunidad activa y el apoyo de la Apache Software Foundation, asegura que Spark se mantenga a la vanguardia de las tecnologías de procesamiento de datos. Su capacidad para adaptarse y evolucionar con las necesidades cambiantes de la industria del Big Data garantiza su relevancia y utilidad a largo plazo.

databricks

IMMUNE
TECHNOLOGY INSTITUTE

Fundación de Databricks

Databricks fue fundada en 2013 por los creadores originales de Apache Spark con el objetivo de facilitar el uso de Spark y acelerar su adopción en la industria. La compañía se estableció para ofrecer una plataforma unificada que integra todos los aspectos del procesamiento de datos y el análisis de Big Data, desde la ingestión de datos hasta el machine learning.

Desde su inicio, Databricks ha buscado simplificar y democratizar el análisis de datos y la ciencia de datos, permitiendo a las organizaciones de todos los tamaños aprovechar el poder de Spark sin necesidad de una compleja infraestructura de datos. Su plataforma en la nube ofrece una solución escalable y eficiente para el procesamiento de Big Data, la analítica avanzada y el machine learning.

Al proporcionar una plataforma que reduce significativamente la complejidad y el tiempo necesario para obtener insights valiosos de los datos, Databricks ha jugado un papel crucial en la transformación digital de numerosas empresas.

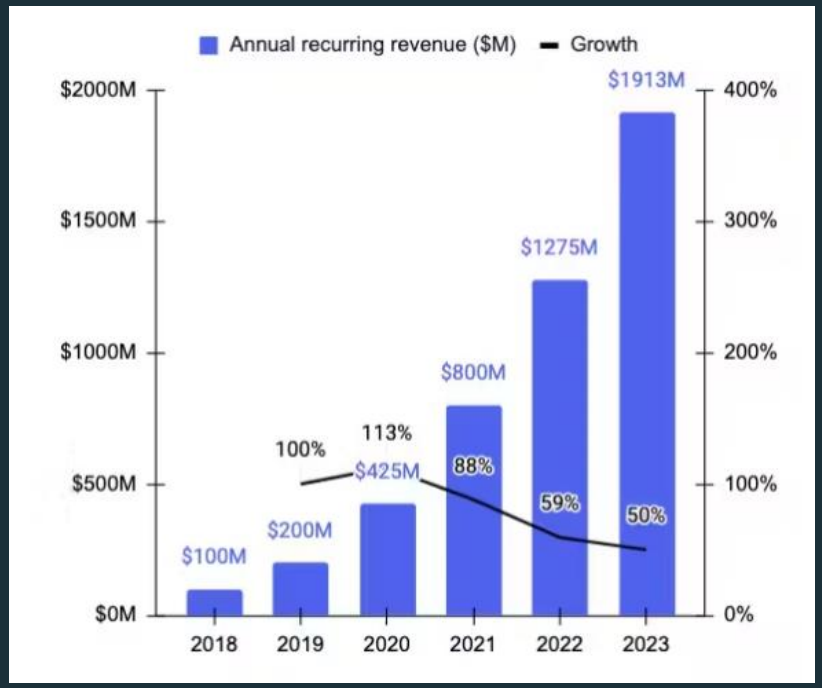
databricks

IMMUNE
TECHNOLOGY INSTITUTE

Crecimiento y Capitalización de Databricks

Databricks ha experimentado un crecimiento exponencial, destacándose en el mercado con su innovadora plataforma. Ha recaudado significativas rondas de financiación, elevando su valoración a decenas de miles de millones de dólares.

Este crecimiento refleja la confianza de los inversores en su tecnología y el valor que aporta a las empresas en su manejo de datos y análisis.



databricks

IMMUNE TECHNOLOGY INSTITUTE

Impacto de Databricks en la Industria

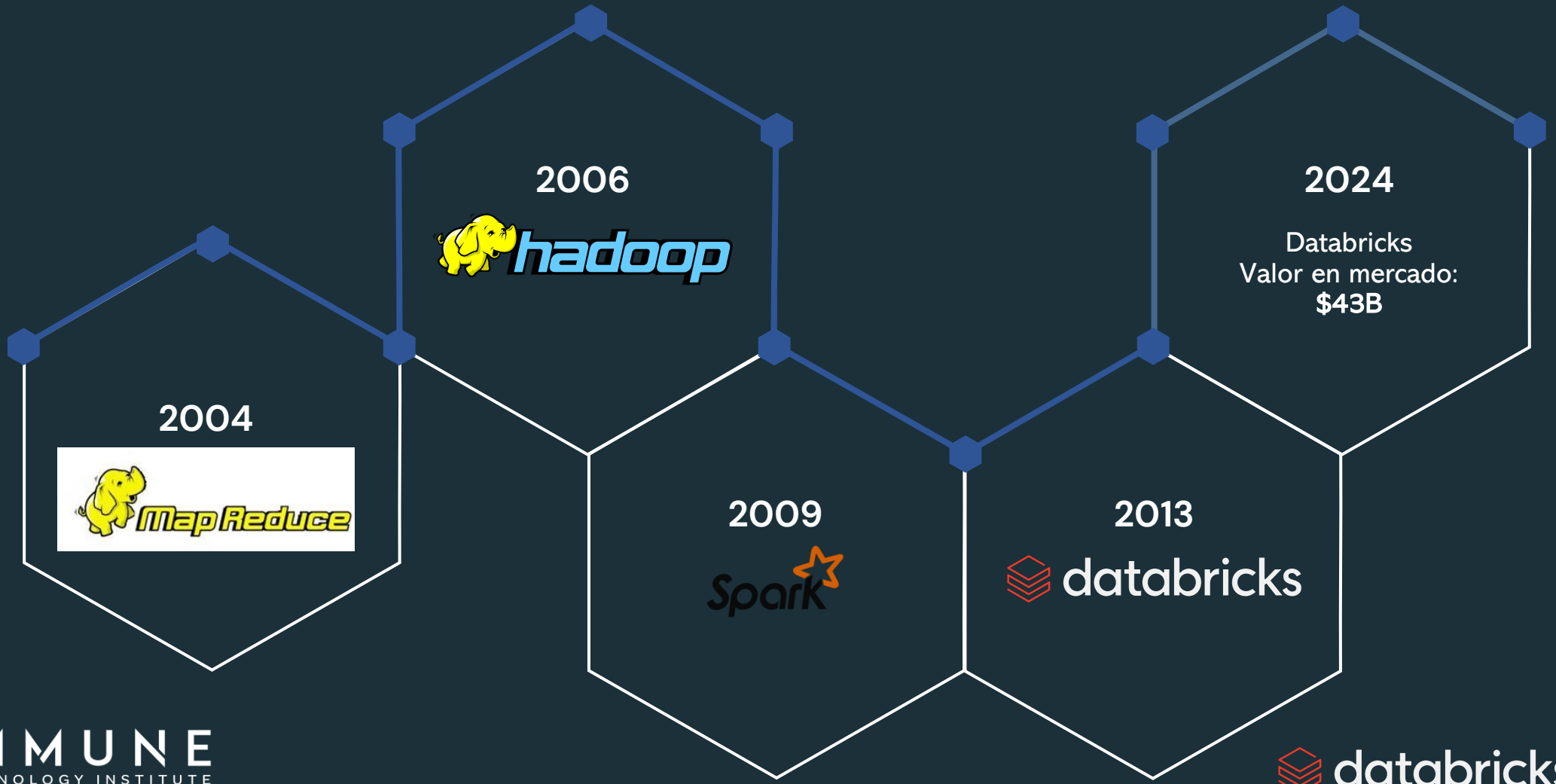
Hoy, Databricks sirve a miles de clientes en todo el mundo, desde startups hasta algunas de las empresas más grandes y respetadas.

Su plataforma ha sido fundamental en la transformación digital de sectores como finanzas, salud, energía y entretenimiento, permitiéndoles innovar y mantenerse competitivos en el mercado global.

databricks

IMMUNE TECHNOLOGY INSTITUTE

Evolución



IMMUNE
TECHNOLOGY INSTITUTE

databricks