**PREWORK**

Sesión 08

**Introducción a NoSQL (Cassandra)**

**🎯 Objetivo**

* Comprender y delimitar el alcance de NoSQL para Big Data
* Introducir al alumno conceptos elementales de Apache Cassandra para la administración de Big Data
* Aprender sobre la historia reciente de este software líder en la industria

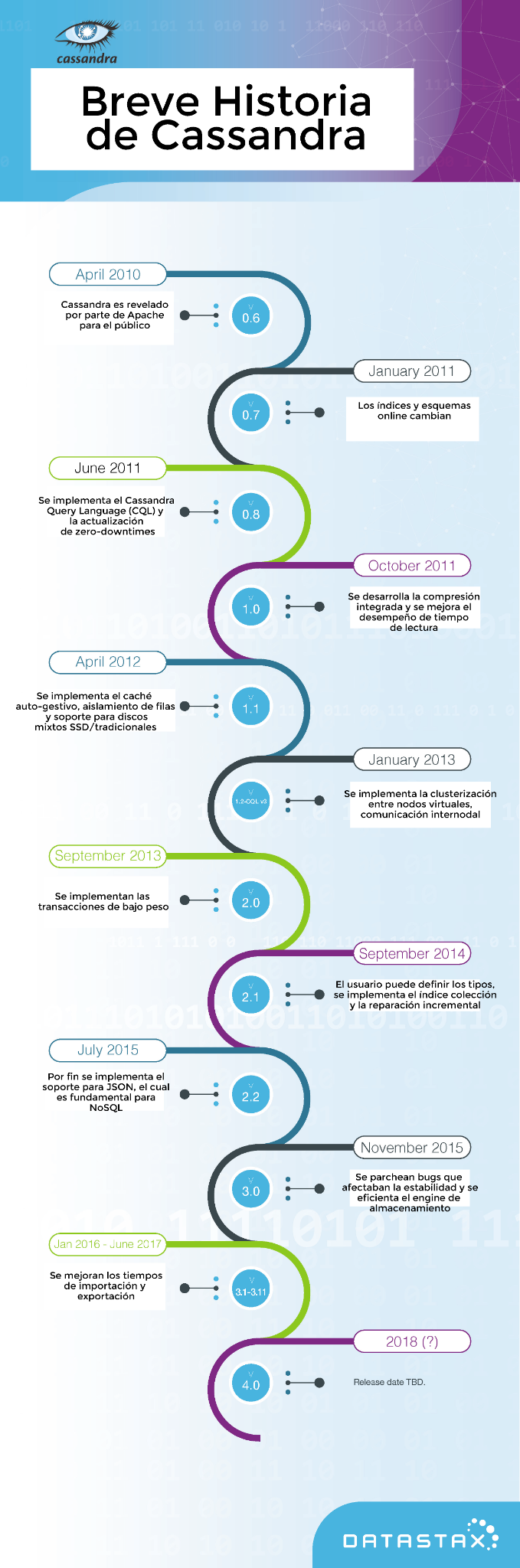
### 🚀 **Desarrollo**

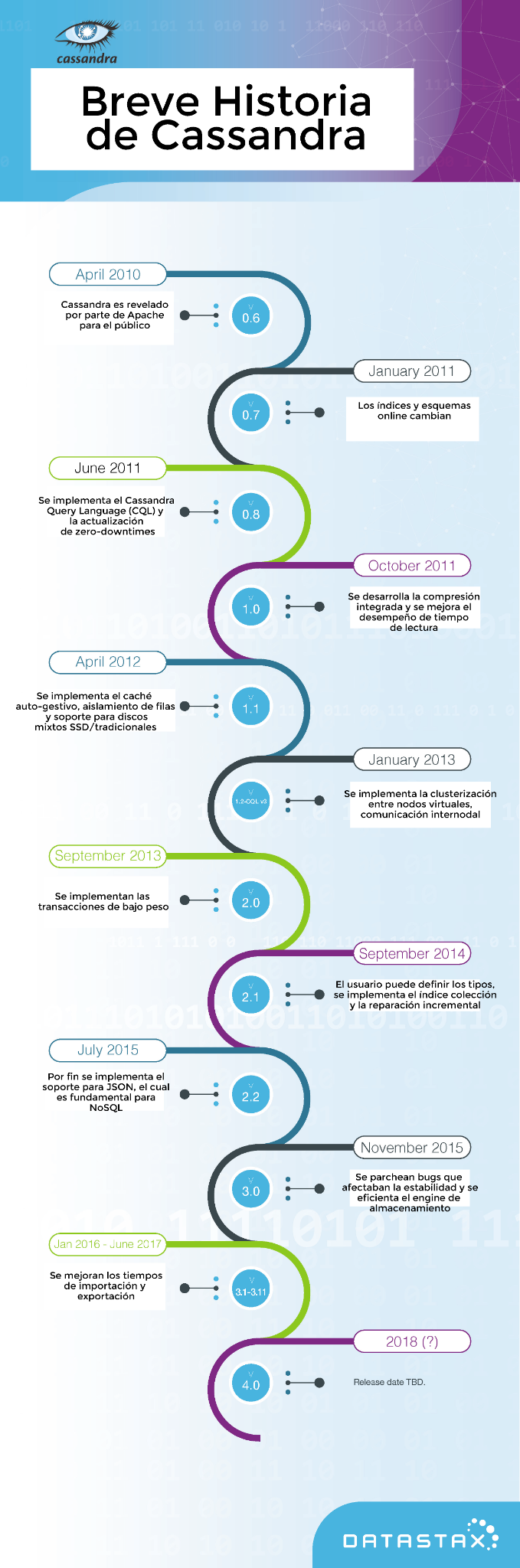
### **Parte 1: ¿Qué es Cassandra?**

Apache Cassandra no es la sacerdotisa troyana de Apolo de la mitología griega que todos conocemos. Apache Cassandra es el único administrador de bases de datos NoSQL distribuido que ofrece disponibilidad permanente, un rendimiento de lectura y escritura increíblemente rápido y una escalabilidad lineal ilimitada que le permite satisfacer las demandas de las aplicaciones modernas exitosas.

Apache Cassandra es de código abierto que comenzó a usarse internamente en Facebook y fue lanzado como un proyecto de código abierto en julio de 2008. Cassandra ofrece disponibilidad continua (tiempo de inactividad cero), alto rendimiento y escalabilidad lineal que las aplicaciones modernas requieren, además de ofrecer simplicidad operativa y replicación sin esfuerzo en centros de datos. Cassandra puede manejar petabytes de información y miles de operaciones simultáneas por segundo, lo que permite a las organizaciones administrar grandes cantidades de datos en entornos de nubes híbridas y múltiples: Cassandra tiene un potencial increíble y hermoso para la industria del Big Data y para algunos informáticos, puede definirse como el futuro de la industria.

A continuación, te presentamos una breve infografía sobre la historia de este importante software:





**¿Cuáles son los requisitos de Hardware para correr Cassandra?**

Cassandra emplea varias técnicas de almacenamiento en caché para permitir operaciones de lectura ultrarrápidas; por lo tanto, más memoria

permite a Cassandra almacenar más datos en caché... más memoria conduciría a un mejor rendimiento. Se recomienda una memoria mínima de 4 GB para entornos de desarrollo y una memoria mínima de 8 GB para entornos de producción. Si nuestro conjunto de datos es más grande, deberíamos considerar actualizar la memoria utilizada por Cassandra. Al igual que la memoria, una mayor cantidad de CPU ayuda a Cassandra a funcionar mejor, ya que Cassandra realiza sus tareas al mismo tiempo. Para hardware físico, se recomiendan servidores de 8 núcleos y para máquinas virtualizadas se recomienda que los ciclos de CPU asignados a las máquinas puedan crecer según la demanda, por ejemplo, algunos proveedores como Rackspace y Amazon utilizan “CPU bursting”. Para los entornos de desarrollo, puede utilizar una máquina de un solo disco; sin embargo, para las máquinas de producción, lo ideal es que haya al menos dos discos. Un disco se utiliza para el registro de confirmación y el otro para almacenar archivos de datos llamados SSTables, de modo que no se produzca contención de lectura y escritura. Cassandra utiliza un archivo de registro de confirmación para hacer que las solicitudes de escritura sean duraderas. Cada solicitud de escritura se escribe primero en este archivo en modo de solo anexión y una representación en memoria de la familia de columnas llamada “memtable”.

**¿Qué ventajas ofrece Apache Cassandra?**

**Open Source:** Las organizaciones modernas de desarrollo de software se han movido abrumadoramente para adoptar tecnologías de código abierto, comenzando con el sistema operativo Linux y progresando hacia la infraestructura para administrar datos. Las tecnologías de código abierto son atractivas debido a su asequibilidad y extensibilidad, así como a la flexibilidad para evitar el bloqueo de proveedores. Las organizaciones que adoptan el código abierto informan una mayor velocidad de innovación y una adopción más rápida.

**Flexibilidad:** El Cassandra Query Language (CQL) es similar a SQL, lo que significa que a la mayoría de los desarrolladores les resultará bastante fácil familiarizarse con él. (Aquí hay una introducción a CQL si necesita ayuda).

**Desempeño de alto rendimiento:** La mayoría de las bases de datos tradicionales cuentan con una arquitectura primaria/secundaria. En estas configuraciones, una única réplica principal realiza operaciones de lectura

y escritura, mientras que las réplicas secundarias solo pueden realizar operaciones de lectura. Las desventajas de esta arquitectura incluyen una mayor latencia, así como mayores costos y menor disponibilidad a escala. En Cassandra, ningún nodo está a cargo de replicar datos en un clúster. En cambio, cada nodo es capaz de realizar todas las operaciones de lectura y escritura. Esto mejora el rendimiento y agrega resistencia a la base de datos.

**Active Everywhere / Zero-Downtime**: dado que cada nodo de Cassandra es capaz de realizar operaciones de lectura y escritura, los datos se replican rápidamente en entornos de computo en la nube. En caso de que falle un nodo, los usuarios se enrutan automáticamente al nodo en buen estado más cercano. Ni siquiera notarán que un nodo se ha desconectado porque las aplicaciones se comportan bien, pues se diseñaron para operar incluso en caso de falla. Como resultado, las aplicaciones siempre están disponibles y los datos siempre están accesibles y nunca se pierden. Además, los servicios de reparación integrados de Cassandra solucionan los problemas inmediatamente después de que ocurren, sin ninguna intervención manual. La productividad ni siquiera necesita verse afectada si los nodos fallan.

**Escalabilidad:** en entornos tradicionales, escalar aplicaciones es un proceso costoso y que requiere mucho tiempo y que generalmente se logra escalando verticalmente con máquinas más caras. Cassandra le permite escalar horizontalmente simplemente agregando más nodos al clúster. Si, por ejemplo, cuatro nodos pueden manejar 200,000 transacciones / segundo, ocho nodos podrán manejar 400,000 transacciones / segundo.

**Replicación perfecta:** las empresas líderes de hoy en día se están moviendo cada vez más hacia implementaciones de múltiples centros de datos, nubes híbridas e incluso múltiples nubes para aprovechar las fortalezas de varias implementaciones sin quedar atrapadas en el ecosistema de un solo proveedor. Sin embargo, aprovechar al máximo los entornos de múltiples nubes comienza con tener una base de datos en la nube subyacente que ofrece: escalabilidad, seguridad, rendimiento y disponibilidad. Por estos motivos, no debería sorprendernos que se espere que el mercado de bases de datos en la nube crezca casi un 65% cada año y alcance los 68.900 millones de dólares en 2022.