Bases de datos NoSQL
Clase 3

Arroyo Joaquin
Belmonte Marina

Universidad Nacional de Rosario
Licenciatura en Ciencias de la Computación
Bases de Datos Avanzadas

12 de junio de 2024

Resúmen

- Estado del Arte
  - The Future of the data storage in Particle Physics and Astronomy
  - Fraud Detection in NoSQL Database Systems using Machine Learning
- Integración de NoSQL en Diversos Contextos
- Conclusiones

Índice

The Future of the data storage in Particle Physics and Astronomy - Introducción

- Desarrollado por Julius Hřivnáč y Julien Peloton, investigadores de la Université Paris-Saclay, en el año 2024.
- Hace incapié en la combinación de bases de datos, en particular "tabulares" (SQL o NoSQL) con Grafos, para el almacenado de datos astronómicos y de experimentos de física de partículas.
- Los experimentos de física de partículas y los telescopios de astronomía almacenan grandes cantidades de datos, principalmente en archivos simples en diversos formatos, con un uso limitado de bases de datos.

The Future of the data storage in Particle Physics and Astronomy - Introducción

Ilustran dicha combinación de bases de datos con el proyecto FINK, uno de los brokers oficiales dentro del Observatorio Rubin y del Zwicky Transient Facility.

Este proyecto utiliza **HBase** y **JanusGraph** para almacenar alertas:

- HBase almacena datos voluminosos.
- JanusGraph almacena información estructural.

The Future of the data storage in Particle Physics and Astronomy - HEP y Grafos

La motivación de proponer una combinación de bases de datos “tabulares” con grafos viene dada por las siguientes razones:

- En la Física de Alta Energía (HEP) el manejo de datos ha dependido durante mucho tiempo de estructuras de datos convencionales.
- Una parte significativa de estos datos exhiben características similares a grafos y carecen de un esquema rígido.
- Estos datos a menudo consisten en entidades conectadas a través de relaciones, lo que los hace poco adecuados para el almacenamiento en bases de datos relacionales.

The Future of the data storage in Particle Physics and Astronomy - HEP y Grafos

Además esto es impulsado también por las propias desventajas que tienen las bases de datos de grafo:

- Inserción Lenta
- Manejo de Memoria Lento
- Creación de Aristas Anárquica
- Esquema Desconocido y Relaciones Caóticas
- Lenguajes de Consulta Avanzados
- etc.

The Future of the data storage in Particle Physics and Astronomy - Solución Híbrida

A la combinación que venimos mencionando la llaman “Solución Híbrida”, la cuál ofrece lo mejor de ambos mundos al combinar las fortalezas de diferentes paradigmas de almacenamiento.

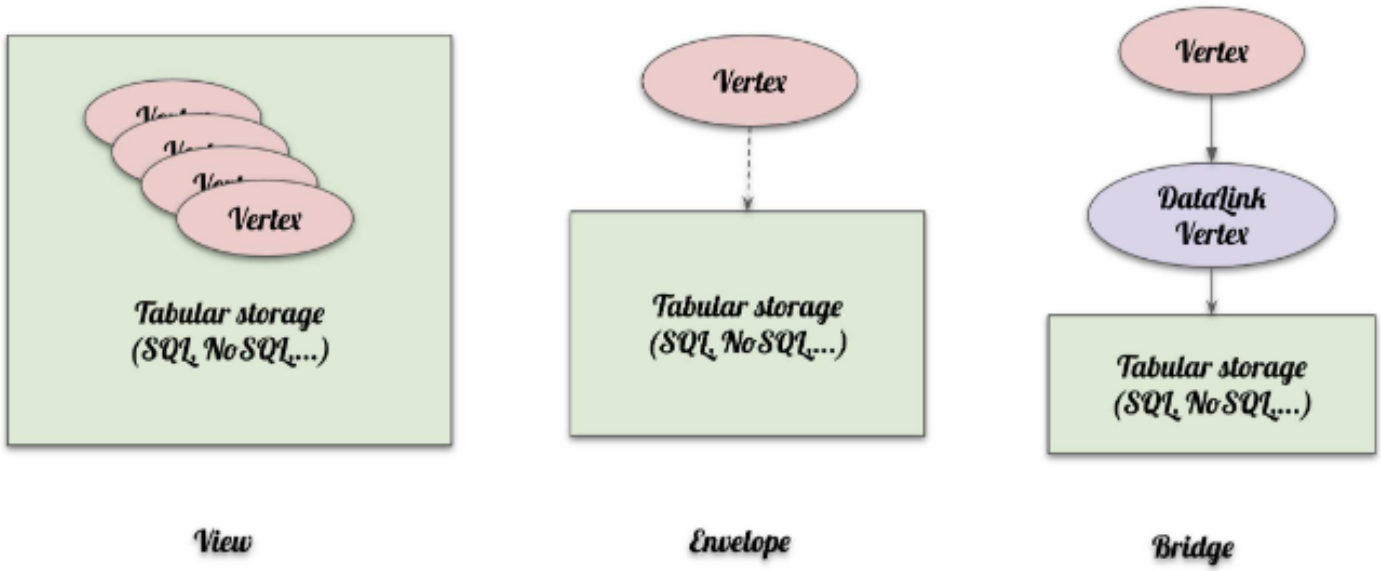
- Esta solución comienza almacenando datos no estructurados o en bruto en bases de datos tabulares.
- Se introduce el concepto de un grafo para expresar y gestionar estructuras de datos persistentes. Este representa relaciones y conexiones complejas dentro de los datos.
- Se conectan estos componentes detrás de una API común.

En esencia:

- Se aprovecha la eficiencia del almacenamiento tipo tabla
- y el poder expresivo de los grafos.

The Future of the data storage in Particle Physics and Astronomy - Solución Híbrida

Tres arquitecturas para conectar bases de datos tabulares y de grafos en una solución híbrida:



The Future of the data storage in Particle Physics and Astronomy - Solución Híbrida

- **Graph View:**
  - Interpreta los datos tabulares existentes como vértices dentro de un grafo.
  - Se agregan aristas adicionales al grafo para expresar relaciones estructurales entre estos vértices.
- **Graph Envelope:**
  - Mejora el concepto de un vértice agregando métodos adicionales para llenarlo desde un almacenamiento tabular externo.
  - Mantener la consistencia entre los datos puede ser complicado.
- **Bridge:**
  - Se crea de un tipo especial de **Vértice de Enlace de Datos** que representa relaciones con datos externos almacenados en cualquier tipo de sistema de almacenamiento.
  - Estos vértices pueden estar conectados a cualquier otro vértice en el grafo, formando efectivamente puentes hacia fuentes de datos externas.

The Future of the data storage in Particle Physics and Astronomy - FINK

- El proyecto FINK es uno de los brokers oficiales dentro del Observatorio Rubin y del Zwicky Transient Facility.
- El observatorio Rubin generará 10 millones de alertas cada noche, lo que equivale a aproximadamente 1 terabyte de datos de alerta con alrededor de 20 terabytes de datos de imagen.
- FINK está recibiendo y analizando los datos del Zwicky Transient Facility, el cuál desde 2019 está enviando en promedio 200,000 alertas por noche, las cuáles se procesan y distribuyen en tiempo real.

The Future of the data storage in Particle Physics and Astronomy - FINK

- Un aspecto esencial de FINK es la presencia de enlaces de datos que conectan los datos en JanusGraph con los datos correspondientes en HBase.
- Estos enlaces de datos sirven como puentes entre los datos estructurados y orientados a grafos y los datos en bruto y tabulares almacenados en HBase.
- Esta combinación forma un sistema robusto de gestión de datos dentro del sistema, lo que permite un almacenamiento, organización y recuperación eficientes de los datos de alerta.

The Future of the data storage in Particle Physics and Astronomy - FINK

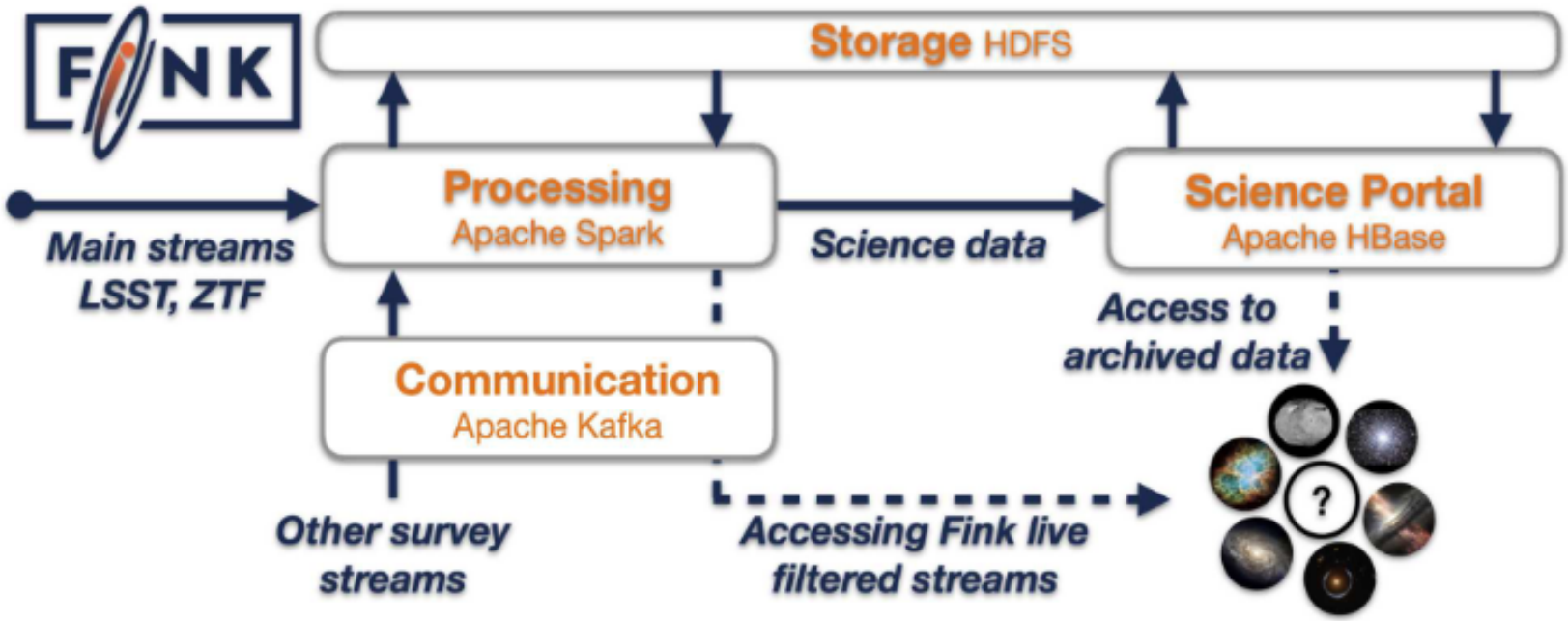


Figura: La arquitectura del broker FINK.

The Future of the data storage in Particle Physics and Astronomy - Conclusiones

- Se destaca la importancia de estructurar los datos de manera efectiva en el contexto de la Física de Altas Energías y las ventajas de utilizar soluciones de almacenamiento híbridas.
- Las soluciones de almacenamiento híbridas combinan la expresividad y flexibilidad de las bases de datos de grafos con el rendimiento y la simplicidad del almacenamiento tabular, proporcionando lo mejor de ambos mundos.
- Las bases de datos híbridas pueden mejorar la forma en que se manejan los datos en experimentos de HEP.

Índice

Fraud Detection in NoSQL Database Systems using Advanced Machine Learning

- Autor: Tamilselvan Arjunan.
- Publicado en: International Journal of Innovative Science and Research Technology.
- Fecha: Marzo 2024.
- Análisis de vulnerabilidades en MongoDB y Cassandra.
- Propuesta de algoritmos de aprendizaje automático.

Fraud Detection in NoSQL Database Systems using Advanced Machine Learning - Problemas de Seguridad en NoSQL

- Esquemas Dinámicos.
- Falta de Control de Acceso.
- Consistencia Eventual.
- Datos Desnormalizados.
- Configuraciones Inseguras por Defecto.

Fraud Detection in NoSQL Database Systems using Advanced Machine Learning - Soluciones Propuestas

- Monitoreo en Tiempo Real.
- Algoritmos de Aprendizaje Automático:
  - Modelos Supervisados.
  - Técnicas No Supervisadas.
  - Métodos en Línea.
- Ingeniería de Características.
- Sistema Híbrido de Detección.
- Aprendizaje Adversarial.
- Implementación y Despliegue.

Fraud Detection in NoSQL Database Systems using Advanced Machine Learning - Conclusiones

- Las técnicas avanzadas de aprendizaje automático mejoran la seguridad.
- Mejorar los mecanismos de control de acceso y autenticación permitirían reducir la superficie de ataque.
- Superar desafíos de integración y necesidad de datos etiquetados.
- Fomentar la colaboración entre equipos de desarrollo y seguridad para asegurar la implementación efectiva de medidas de protección.

Integración de NoSQL en Diversos Contextos

- **Bases de Datos Espacio-Temporales**: MongoDB ofrece soporte nativo para índices geoespaciales y consultas de rango temporal.
- **Toma de Decisiones**: Couchbase y MongoDB permiten realizar análisis multidimensionales directamente sobre los datos almacenados.
- **Espacios Métricos**: Las bases NoSQL pueden implementar estructuras de indexación específicas para manejar búsquedas en espacios métricos.
- **Datos en la Web**: Las bases de datos documentales y las bases de datos de grafos, son ideales para almacenar datos semiestructurados y gráficos RDF, proporcionando la flexibilidad necesaria para modelar datos semánticos.

Conclusiones

- Las bases de datos NoSQL han ganado gran popularidad debido a su flexibilidad para manejar datos no estructurados o semiestructurados, su escalabilidad para adaptarse a grandes volúmenes de información y su facilidad de uso en entornos distribuidos.
- No vienen a reemplazar por completo a las bases de datos relacionales. Estas últimas siguen siendo la mejor opción para gestionar datos estructurados que requieren relaciones complejas y consultas ACID.
- Existen propuestas firmes que alientan a la combinación de bases de datos NoSQL y relacionales. Esto puede ser un enfoque superador que permita aprovechar las fortalezas de cada una.
- Existen una gran variedad de implementaciones, cada una con sus propias características y enfoque, lo que permite elegir la solución más adecuada para cada caso de uso específico.

Dudas?

Referencias

1 Hřivnáč, J., Peloton, J.: Multidatabase the future of the data storage in particle physics and astronomy. EPJ Web of Conferences 295 (2024) doi.org/10.1051/epjconf/202429501039

2 LSST Science Book, Version 2.0 (2009)

3 Fink Project, fink-broker.org

4 Zwicky transient facility (ztf), www.ztf.caltech.edu

5 Janusgraph, janusgraph.org

6 Moller, A., Peloton, J., Ishida, E.E.O., Arnault, C., Bachelet, e.a.: fink, a new generation of broker for the LSST community. Monthly Notices of the Royal Astronomical Society 501(3), 3272–3288 (2020), doi.org/10.1093/mnras/staa3602

7 Arjunan, T.: Fraud Detection in NoSQL Database Systems using Advanced Machine Learning (2024) doi.org/10.38124/ijisrt/IJISRT24MAR127