

Integrating NoSQL into VO to support Big Data Challenges

José Antonio Magro Cortés

TFM - Máster en Métodos y Técnicas Avanzadas en Física

Contenido

- 1 Objetivo
- 2 Motivación
- 3 Big Data
 - Definición
 - Problemas
- 4 Observatorio Virtual
- 5 NoSQL
 - Definición
 - ¿Qué es un documento?
 - MongoDB
 - Casos de éxito
- 6 NoSQL en el VO
- 7 Conclusiones y trabajo futuro

Objetivo

Realizar un estudio sobre la problemática de *Big Data* en astronomía y proponer tecnologías alternativas a las existentes, dentro del Observatorio Virtual.

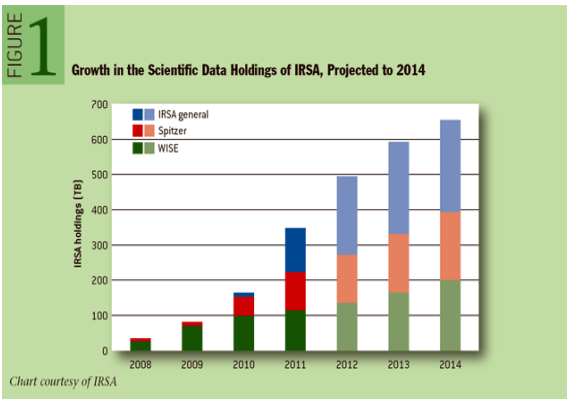
Motivación

- A raíz del curso *Archivos Astronómicos: El Observatorio Virtual* (MTAF):
 - Problemas en la gestión de grandes cantidades de datos
 - Inadecuación de:
 - DBMS
 - Formatos
 - Tecnologías

Definición

Conjuntos de datos tan grandes y complejos que dificultan su procesamiento mediante los sistemas actuales y tradicionales de gestión de bases de datos. Los problemas principales son la captura, clasificación, sistematización, almacenamiento, búsqueda, compartición, transferencia, análisis y visualización.

Problemas



Observatorio Virtual

- FITS
- TAP
- OpenCADC

Definición

NoSQL = *Not Only SQL*

Tipos: orientadas a **documentos**, grafos, objetos, tabulares, etc.

Ventajas

- Escalabilidad
- Modelos de datos flexibles
- Bajo coste

Inconvenientes

- Relativamente reciente
- Soporte
- Pocos usuarios

¿Qué es un documento?

Relational Model



Document Model



MongoDB

- Orientada a documentos
- Drivers para varios lenguajes
- Soporte completo para índices
- Fácil instalación y puesta en marcha
- Balanceo de carga
- Soporta MapReduce
- Conexión con Hadoop
- Tecnología actual para problemas actuales

Casos de éxito

- CMS en el LHC: 10 PB de datos cada año.
- ATLAS Workload Management System.
- Medida de niveles de radiación en Seattle.

NoSQL en el VO

- Almacén FITS
- MapReduce para registros
- Conexión OpenCADC y NoSQL para ALMA

Conclusiones y trabajo futuro

Conclusiones

- NoSQL más eficiente para algunos problemas
- Menor coste para análisis y diseño
- Facilidad para incorporar frameworks de VO a NoSQL

Trabajo futuro

- Adaptar OpenCADC a NoSQL
- Usar métricas formales de diseño
- Benchmarks para medir rendimiento