

Integrating NoSQL into VO to support Big Data Challenges

José Antonio Magro Cortés

TFM - Máster en Métodos y Técnicas Avanzadas en Física

Contenido

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):
 - 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

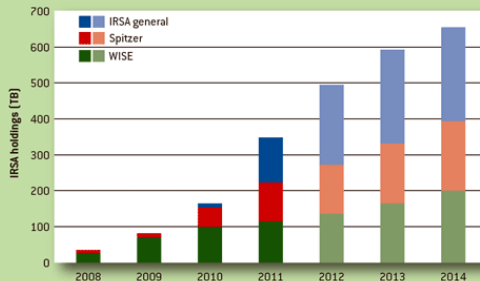
FIGURE 1**Growth in the Scientific Data Holdings of IRSA, Projected to 2014**

Chart courtesy of IRSA

Observatorio Virtual

- FITS
- TAP
- OpenCADC

NoSQL

NoSQL = *Not Only SQL*

Ventajas

- Escalabilidad
- Modelos de datos flexibles
- Bajo coste

Inconvenientes

- Relativamente reciente
- Soporte
- Pocos usuarios

NoSQL: Casos de éxito

- CMS en el LHC
- ATLAS Workload Management System
- Medida de niveles de radiación en Seattle

NoSQL: MongoDB

- Código abierto
- Soporte completo para índices
- Fácil instalación y puesta en marcha
- Balanceo de carga
- Soporta MapReduce
- Conexión con Hadoop

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