

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 (I)
 - Problemas (II)
 - Problemas (III)
 - Solución
- 4 NoSQL
 - Definición
 - ¿Qué es un documento?
 - MongoDB
 - Casos de éxito
- 5 NoSQL en el VO
 - Problemas en el VO
 - Solución: NoSQL en el VO
- 6 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 volúmenes de datos en el VO.
 - Inadecuación de:
 - DBMS.
 - Formatos.
 - Tecnologías.

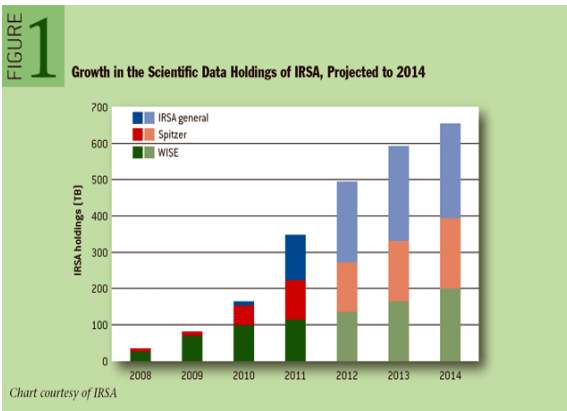
Definición

big data *noun*: ¹

referring to technologies and initiatives that involve data that is too diverse, fast-changing or massive for conventional technologies, skills and infrastructure to address efficiently.

¹10gen Big Data Whitepaper

Problemas (I)



Problemas (II)

- ALMA: 200 TB/año en el ALMA Archive.
- SKA: *ultimate Big Data challenge*, 10-500 TB/s.
- MWA: 3 PB/año.
- LSST: 30 TB/día.

Problemas (III)

Bases de datos relacionales:

- Ineficientes para el procesamiento distribuido.
- Limitaciones en la velocidad de las consultas.
- Falta de escalabilidad.

Solución

MongoDB: Base de datos NoSQL orientada a documentos.



Definición

NoSQL = *Not Only SQL*

Tipos de BD NoSQL: 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.

Problemas en el VO

- FITS
 - Almacenamiento en archivos.
 - Varias convenciones: IDI, SD, MB, OI.
 - Múltiples combinaciones clave-valor.
- TAP: sólo lenguajes relacionales (ADQL/PQL).
- OpenCADC: diseñado para RDBMS.

Solución: NoSQL en el VO

- Almacén FITS en documentos MongoDB.
- MapReduce para grabar registros.
- Conexión OpenCADC y NoSQL para ALMA: Java y driver MongoDB.

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