LATIN AMERICA™ GEOSPATIAL FORUM

Arquitectura para gestionar Big Data en Sistemas de Recomendaciones sensibles al contexto

Dra. Tatiana Delgado Fernández tdelgado@ind.cujae.edu.cu

Instituto Superior Politécnico José Antonio Echeverría, CUAJE, Cuba

Dr. Guillermo González Suárez guille@geomix.geocuba.cu

Dr. José Luis Capote Fernández capote@geomix.geocuba.cu

Dr. Rafael Cruz Iglesias rcruz@geomix.geocuba.cu

Agencia GeoMix, Empresa GeoSI, Cuba

cujae

Contenido

PARTE I Sistema de Recomendaciones sensible al contexto basado en ontologías

1. Problema con el uso de los servicios de datos geográficos de la IDERC
2. Pregunta de Investigación: ¿Cómo desarrollar servicios de Información Geográfica Ubicua para ciudadanos a partir de la IDERC?
3. Conceptos base: Sistemas de Recomendaciones y la Información Geográfica Ubicua
4. Arquitectura e implementación de CARS sobre la IDERC

PARTE II Big Data Management para Sistema de Recomendaciones sensible al contexto y basado en ontologías

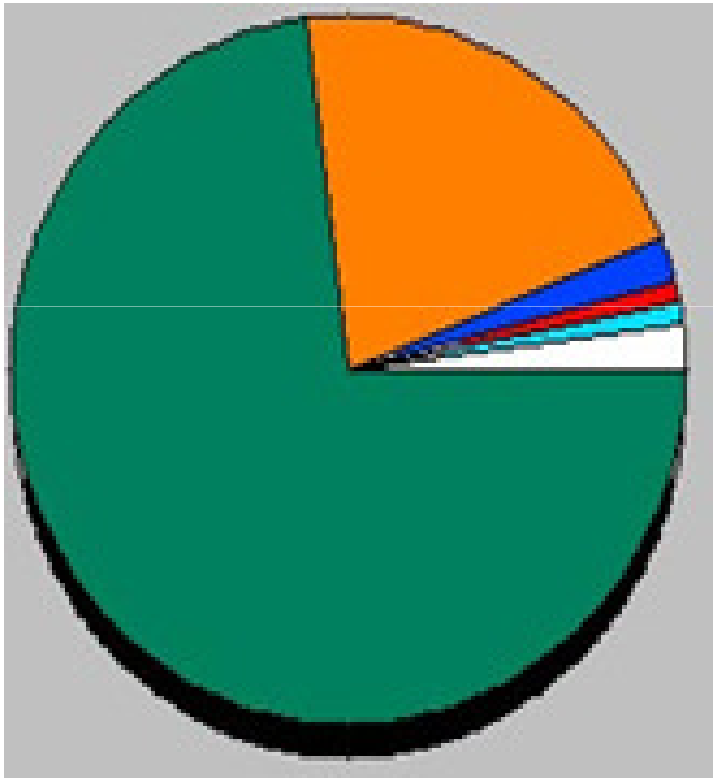
1. Problema de escalabilidad
2. Conceptos base: Big Data
3. Pregunta de Investigación : ¿Cómo escalar los CARS sobre la IDERC en contextos de Big Data?
4. Arquitectura para gestionar Big Data en un CARS soportada en IDEs

FINAL - Trabajo Futuro y Conclusiones

PARTE I

Sistema de Recomendaciones sensible al contexto y basado en ontologías

Uso de la IDE en Cuba (IDERC) (abril 2013)



75% accesos de Cuba

Provenientes de aplicaciones empresariales y el GeoPortal

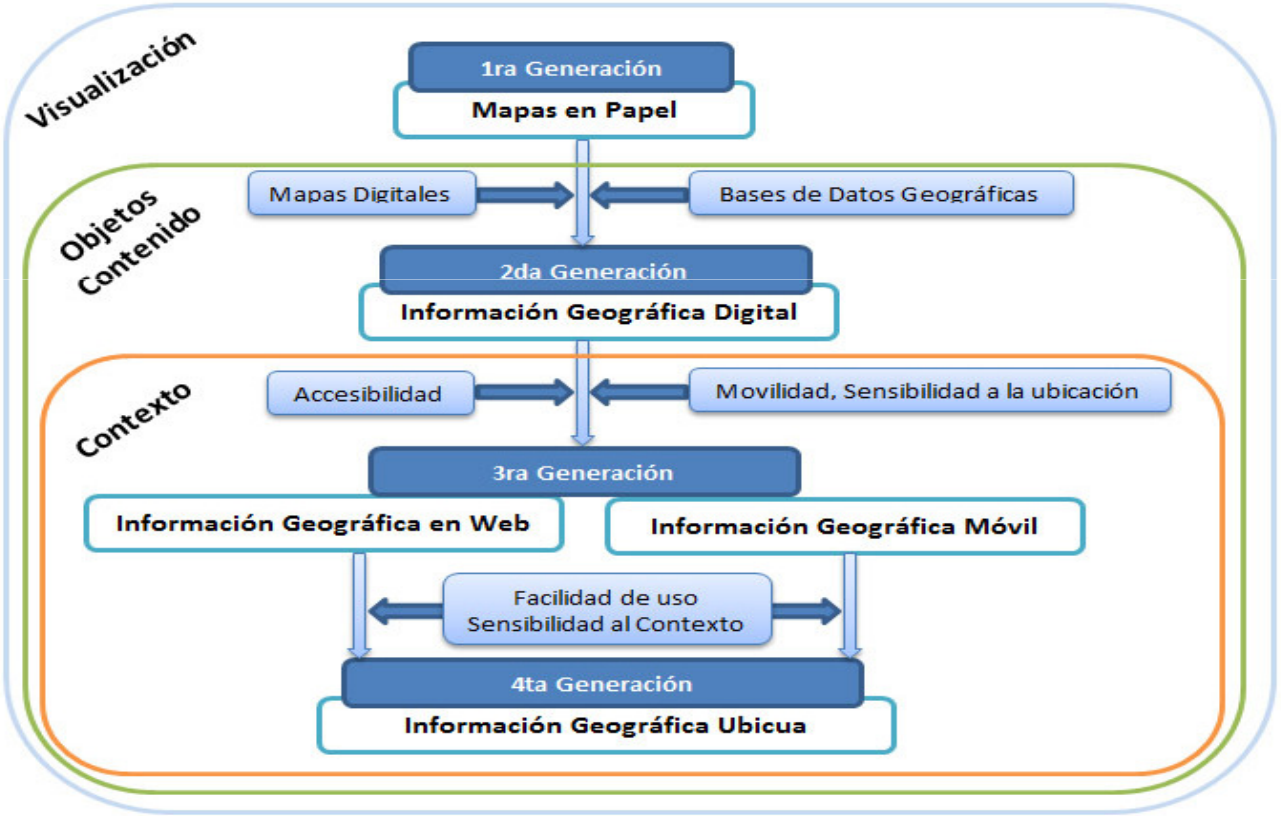
0% aplicaciones orientadas al usuario

Pregunta de Investigación No. 1

¿Cómo desarrollar servicios de Información Geográfica Ubicua para ciudadanos a partir de la IDERC?

Información Geográfica Ubicua

La información geográfica ubicua, es proporcionada en cualquier momento y lugar a usuarios o sistemas utilizando dispositivos de comunicación. Un aspecto crítico de la UBGI es que la información proporcionada se base en el contexto del usuario.

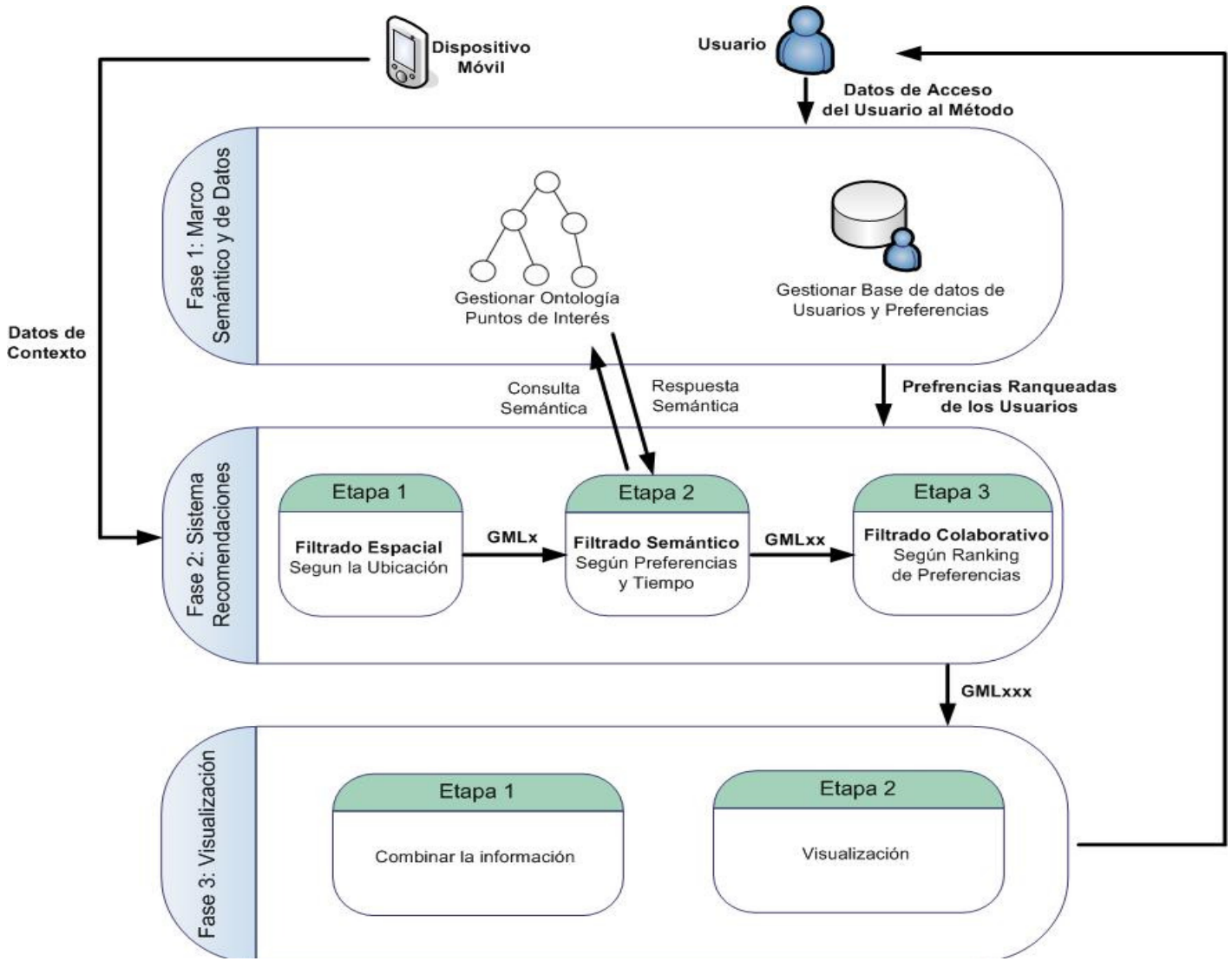


Hong, Sang-Ki. *Ubiquitous Geographic Information (UBGI) and address standards. ISO Workshop on address standards: Considering the issues related to an international address standard.* Copenhagen, Denmark: s.n., 2008.

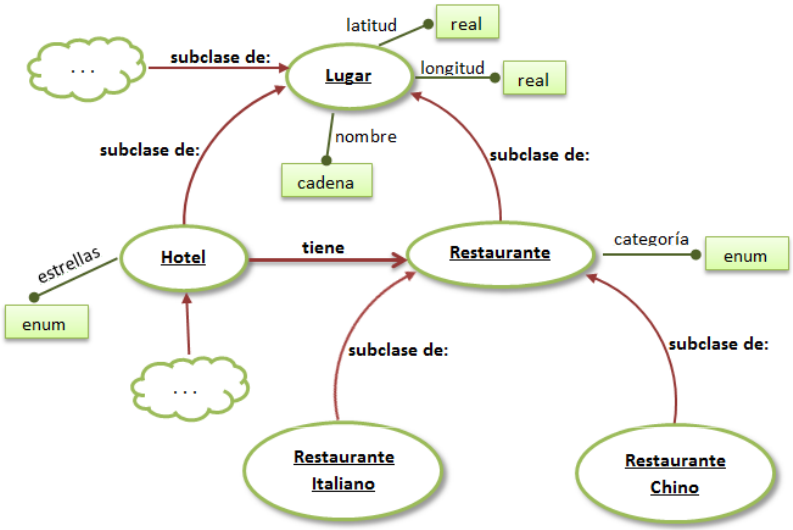
Sistemas de recomendaciones en la Web

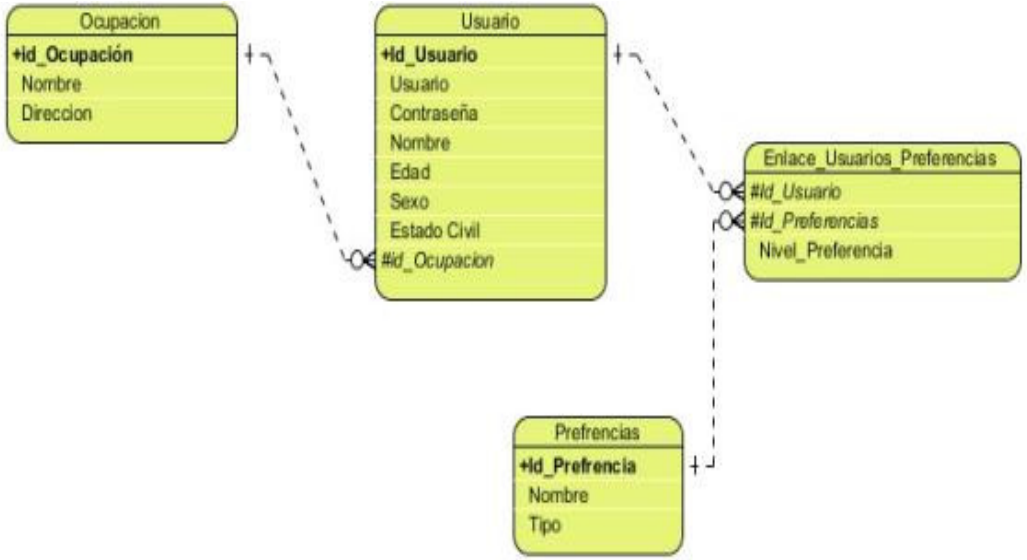
Cualquier sistema que produce recomendaciones individuales como salida, o que tiene el efecto de guiar al usuario de un modo personalizado a objetos útiles y/o interesantes dentro de un gran espacio de posibles opciones.

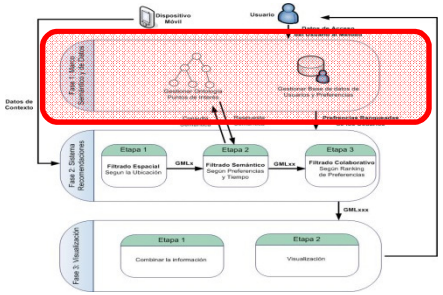
Arquitectura de un Sistema de recomendaciones sensible al contexto basado en ontologías



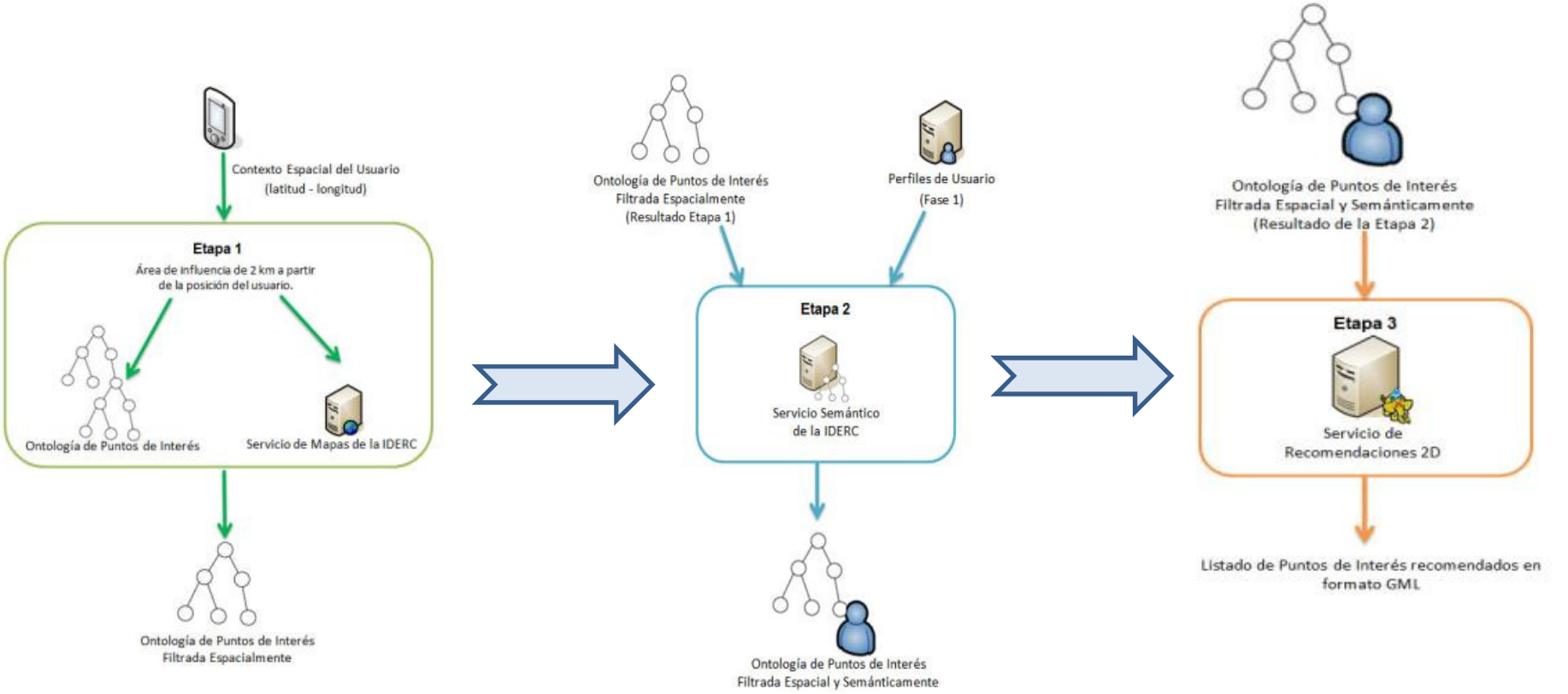
Fase 1.
Marco de datos y semántico

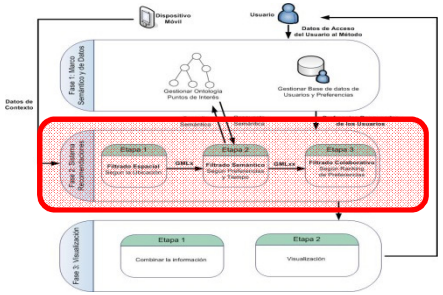






Fase 2.
Pre-filtrado espacio-semántico y filtrado colaborativo





Datos Iniciales
U x P x C x R

Datos Contextualizados
U x P x R

Motor Recomendaciones
U x P ⟶ R

Elementos Recomendados
p1, p2, p3.... pn

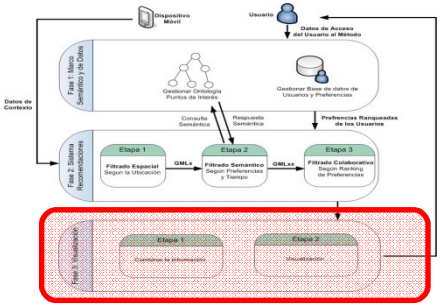
Filtrado Espacial – Semántico (C)

Fase 3. Visualización

Variante 1: Datos de las recomendaciones visualizadas en un formulario

Formularios de Visualización

Variante 2: Datos de las recomendaciones visualizadas en un mapa



Evaluación

Datos:

- Ontología de destinos turísticos.
- Perfiles de usuario.
- Base datos espaciales.

Universo de Lugares: 2 946 destinos de todo el país.
Universo de Usuarios: 60 usuarios diferentes; 36 hombres y 24 mujeres.
Universo de Recomendaciones: 176 760 lugares con su escala de preferencia.

Métricas de evaluación

Medida de precisión de la predicción:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}$$

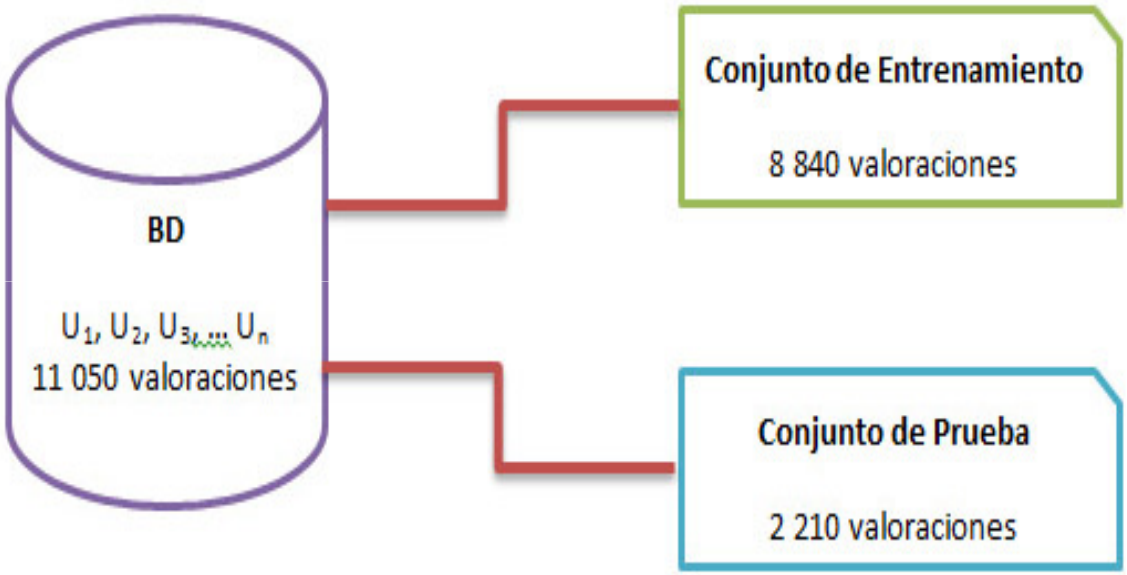
donde:
MAE: Error medio absoluto
p_i: es el valor de la recomendación calculada por el método.
r_i: es el valor que el usuario ha expresado de su preferencia por el elemento i.
N: es la cantidad de elementos del conjunto.

Medida de precisión de la clasificación:

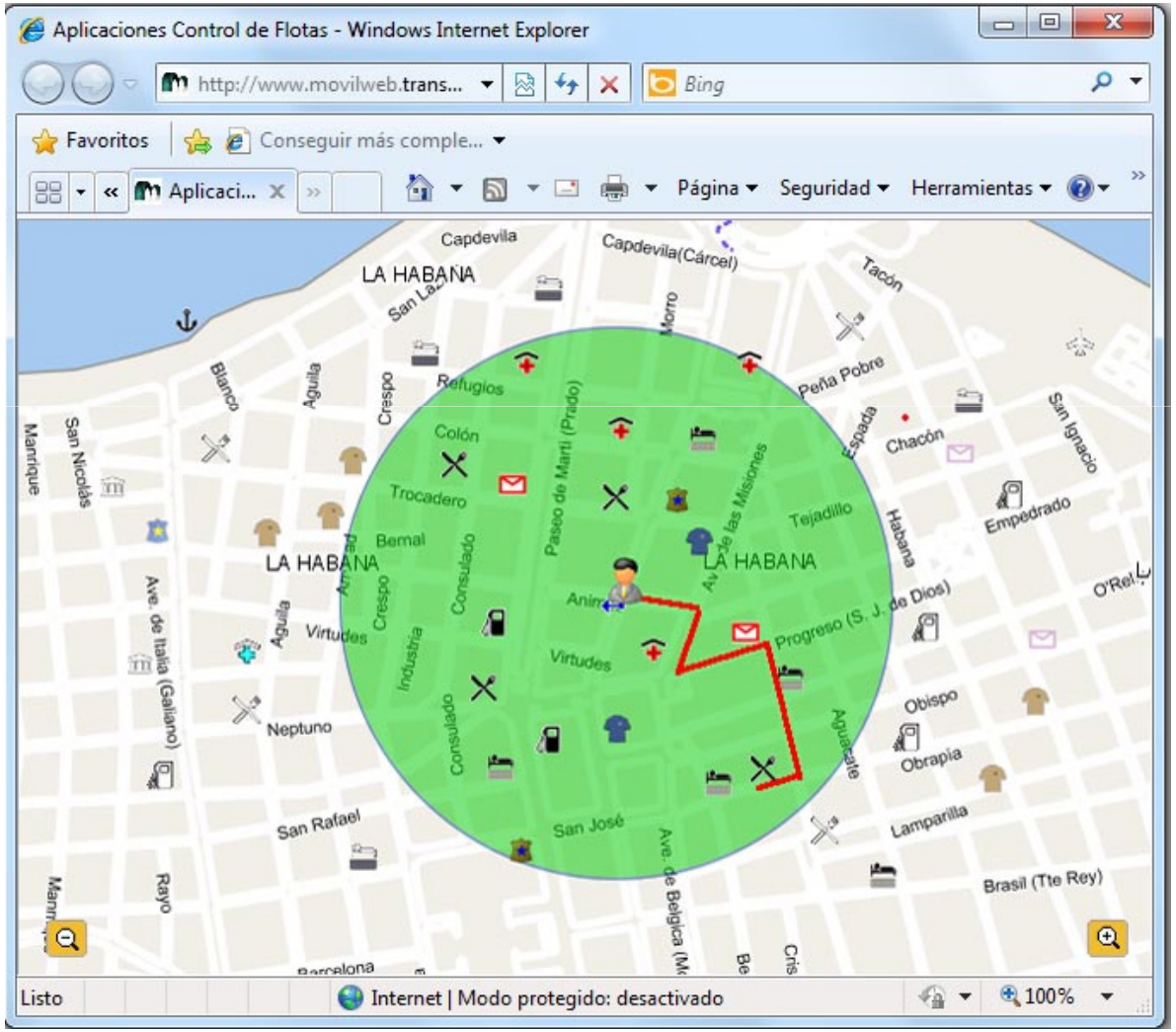
$$Precisión = \frac{N_{rs}}{N_s} \qquad Recuerdo = \frac{N_{rs}}{N_r}$$

donde:
N_s es el número total de elementos seleccionados por el sistema.
N_{rs} es el número de elementos relevantes seleccionados por el sistema.
N_r es el número de elementos que el usuario ha clasificado como relevantes.

Preparación del escenario



Visualización ejemplo evaluación



PARTE II

Usando Big Data en Sistemas de Recomendaciones sensibles al contexto y basados en ontologías

Problemas de escalabilidad para conjuntos de datos mayores de 10 millones

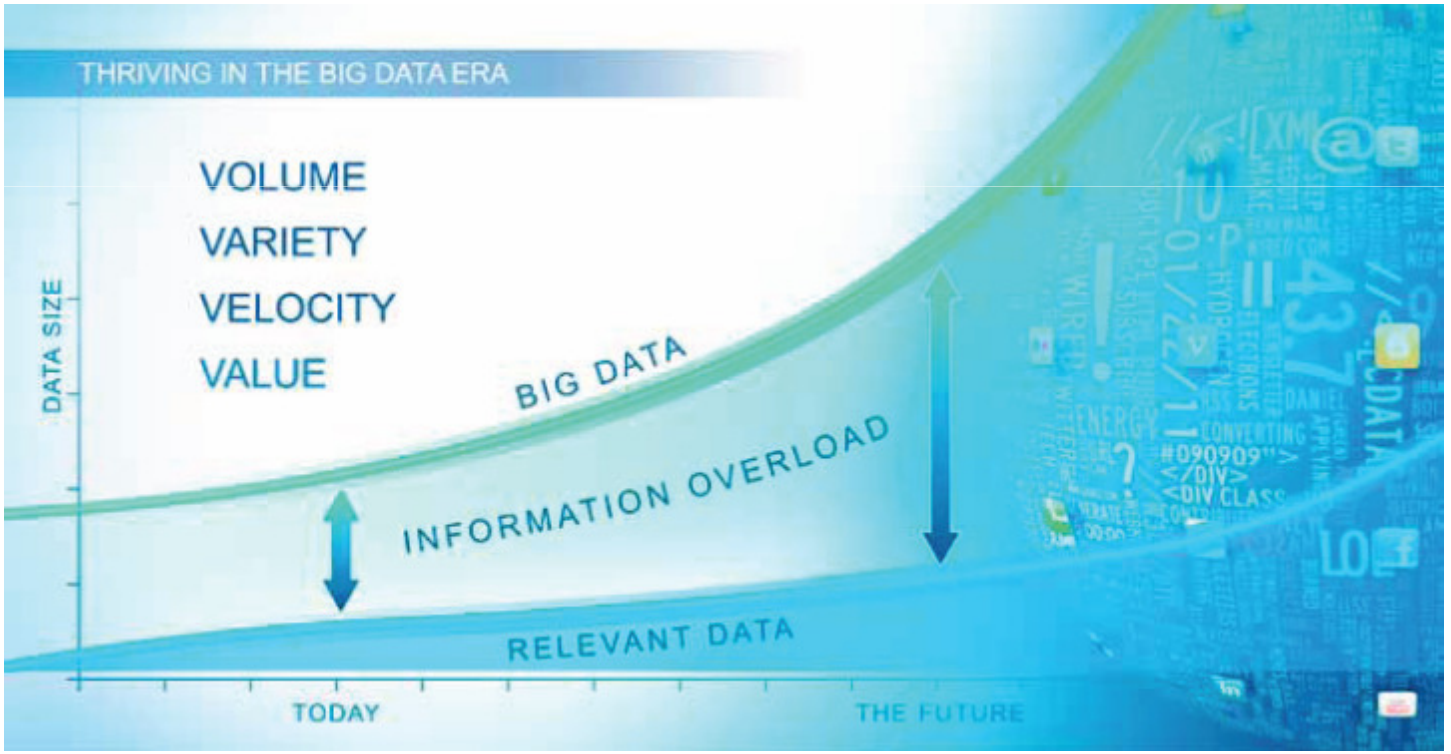
Biblioteca Mahout (Apache Software)

Tamaño de conjunto de datos (usuarios)	Desempeño de Mahout
< 100 000	Más bien lentpo
From 100 000 to 1 milllion	Comienza a ser una buena opción
From 1 to 10 million	Excelente desempeño de Mahout

Owen, S. Anil, R. Dunning, T. and Friedman, E. *Mahout in Action.* New York : Manning Publications Co., ISBN: 9781935182689, 2012.

Big Data

Conjuntos de datos difíciles de manejar que requieren técnicas para agregar, manipular, analizar y visualizarlos.



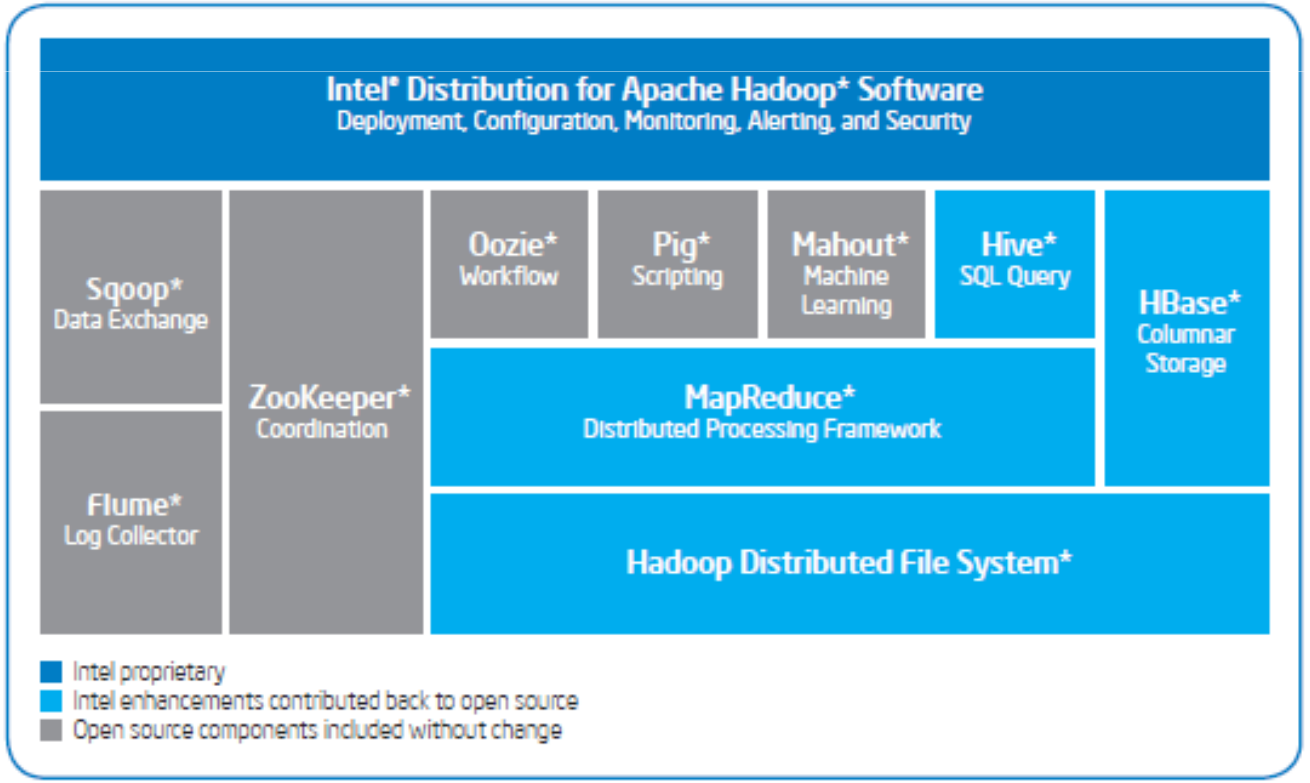
Mark Troester, 2012, SAS, White paper “Big Data Meets Big Data Analytics”

Big Data

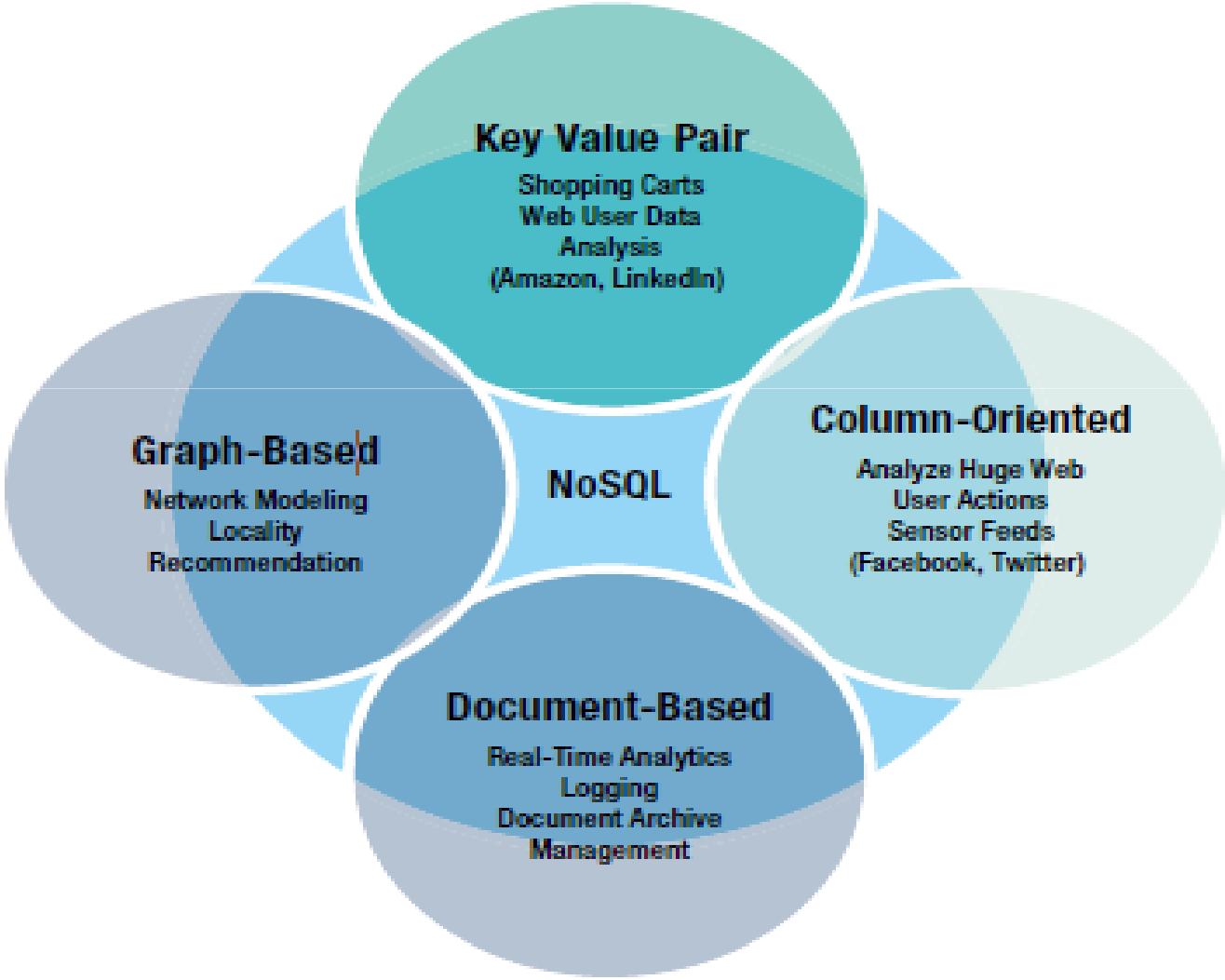
Volume	Velocity	Variety
CARS can generate TB-PB of preferences considering millions of users and millions of Points of Interest.	Social media data streams produce a large influx of opinions valuable to CARS. GPS data need to be analyzed in real-time in CARS	Social media data streams, SMSs, GPS data and RDF (Linked Data) are examples of non-conventional data involved in Spatial CARS

Volumen y Velocidad

- Procesamiento paralelo distribuido de grandes conjuntos de datos a través de clusters de computadoras

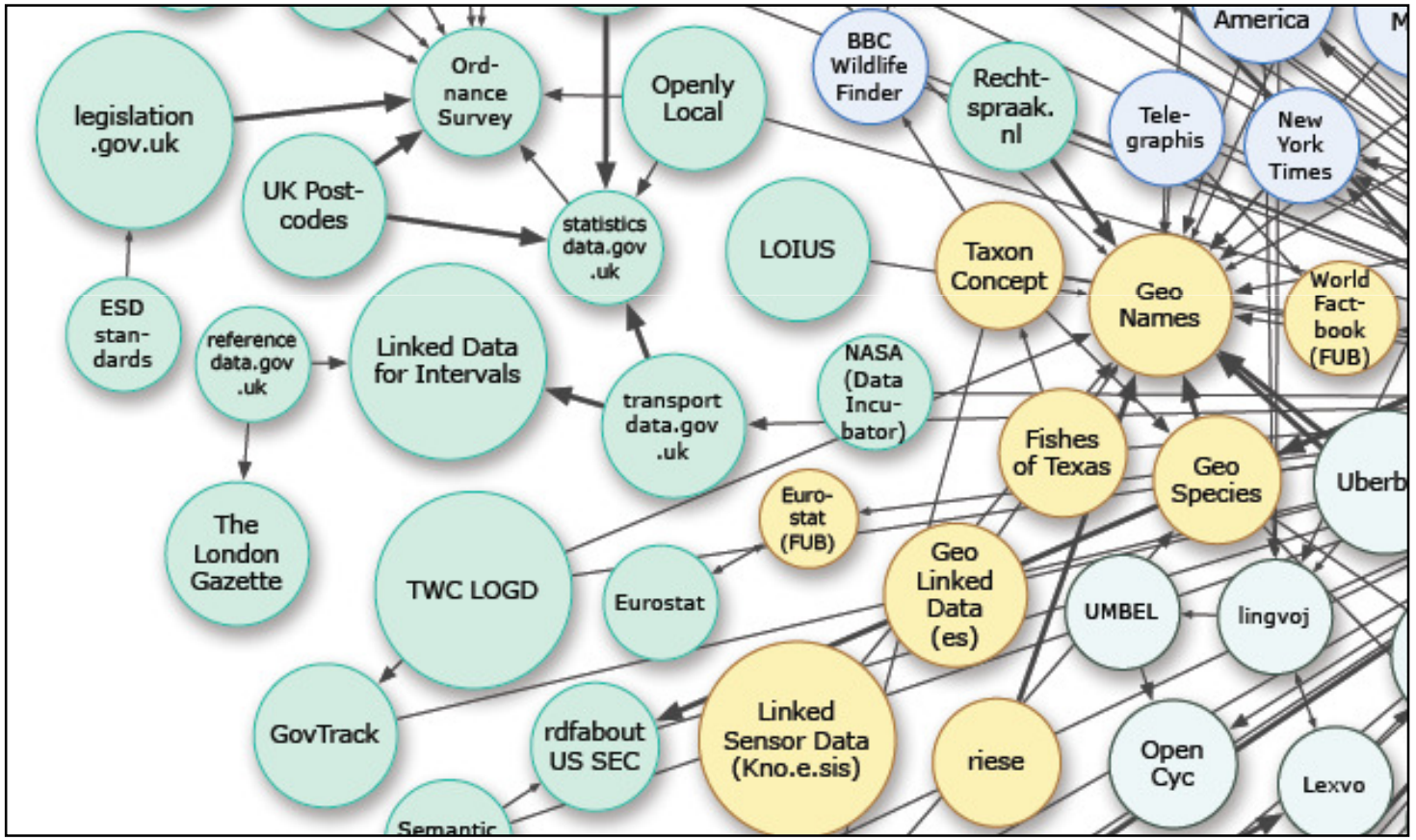


Variedad - NoSQL



Escalar con fuentes de Datos Genéricos de Lugares

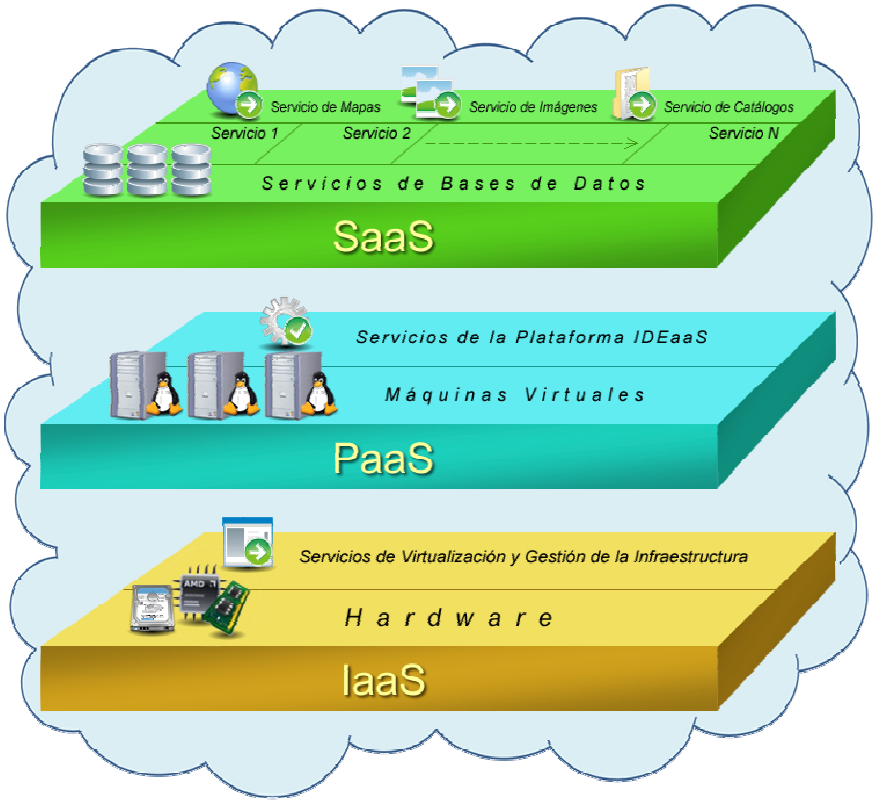
ej. GeoNames (Open Linked Data)



Velocidad:
Hadoop en la Nube Privada de la IDERC

Hadoop puede ser lanzado y correr sobre una nube privada

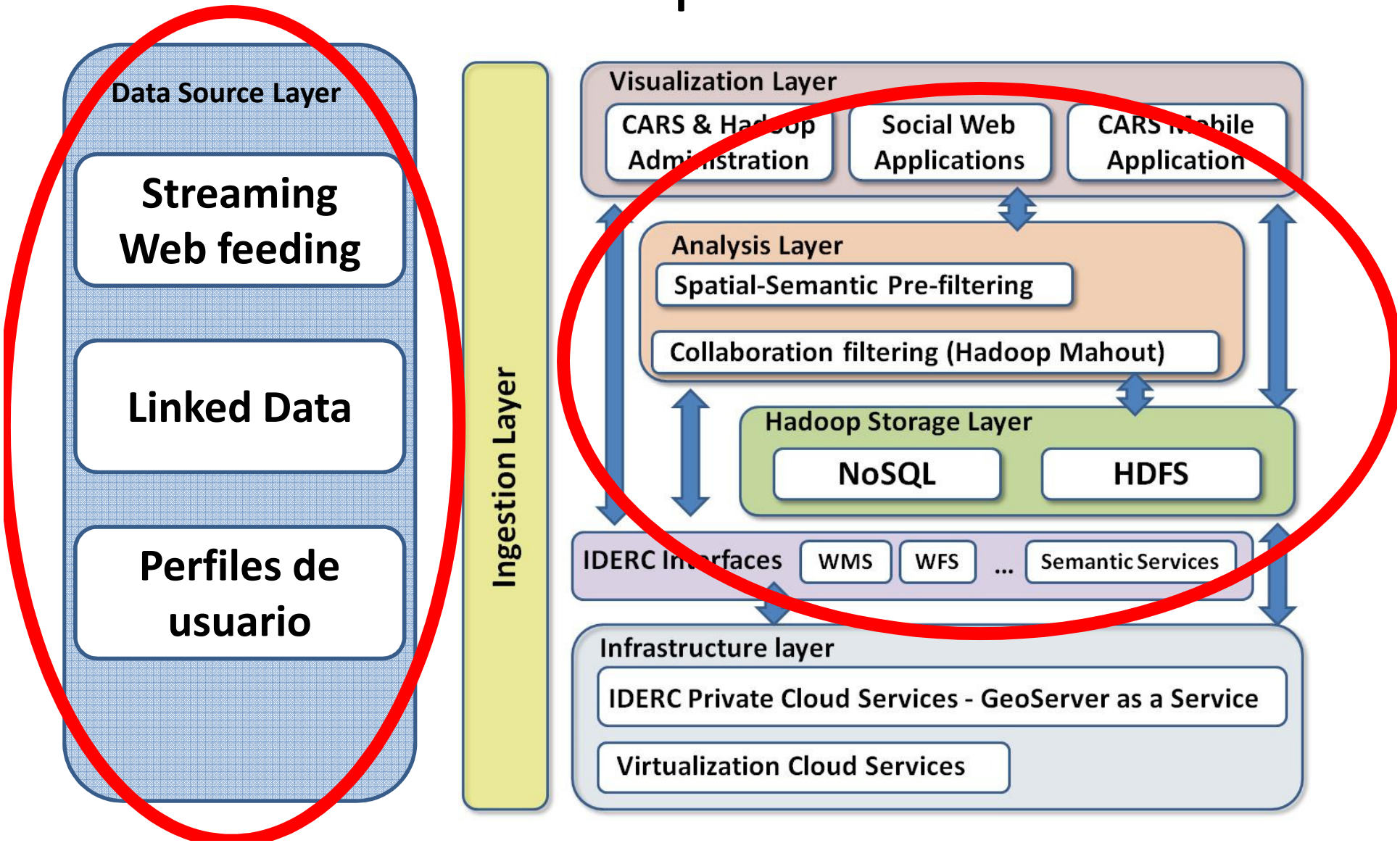
Nube Privada para seguridad de datos y control de acceso u mejor visibilidad y control de la infraestructura, así como visibilidad y control de la infraestructura de Haddop



Hadoop en la Nube

1. **Hadoop corre mejor sobre servidores físicos.** Hadoop comprende un nodo maestro llamado nodo Nombre y múltiples nodos hijos, llamados nodos de datos Estos nodos de datos están en servidores físicos.
2. **Hadoop es "Rack Aware"** – Nodos de datos de Hadoop (servers) son instalados en RACK. Cada RACK contiene muchos servidores.
3. **"Rack awareness"** significa que el nodo nombre conoce dónde cada servidor de nodo de datos está en el RACK

Arquitectura para gestionar Big Data en un CARS soportado en IDEs



Trabajo Futuro

- Implementar la arquitectura propuesta
  - Montar un proyecto piloto en la Universidad para probar las tecnologías de Big Data para cada building block
  - Integrar en una solución el CARS basado en Big Data
- Continuar experimentando con Big Data, en particular para la carga y análisis de información de sensores

Conclusions

- Los CARS basados en IDEs pueden impactar positivamente en la habilitación espacial del ciudadano
- Big Data es una opción recurrente para escalar un CARS
- La arquitectura extendida del Sistema de Recomendaciones usando *Big Data Management*
- El despliegue en la nube de Hadoop mejora la efectividad de la misma.