

Clúster de Investigación Aplicada con enfoque en el Area HPC

Situación actual de proyectos

Facultad Politécnica – UNA

Ing. Julio Paciello

juliopaciello@gmail.com



Contenidos

- **Clúster de Investigación Aplicada**
- Proyectos HPC
- Clúster Hadoop para tecnologías de BI
- Una nube privada para la Administración Pública
- Laboratorios virtuales - *Education as a Service*
- Conclusiones
- Invitación a ERTIC 2013

Clúster de Investigación Aplicada

- **Docentes**
- **Investigadores**
- **Profesionales**
- **Funcionarios Administrativos**
- **Gerentes de proyecto**
- **Directivos**
- **Estudiantes universitarios**



Clúster de Investigación Aplicada

Proyectos actuales

Facultad de Ciencias Médicas



Administración Nacional De Electricidad



Investigación de tecnologías
de BI para Data Warehouse
Sector Privado

Clúster de Investigación Aplicada

- **Afianzar la relación**
 - Universidad-Empresa (pública o privada)
 - Universidad-Universidad (internacionalización)
- **Figura Institucional de la FPUNA para estos proyectos**
 - Generación de Conocimiento
 - Multidisciplinario
 - Profesores con experiencia profesional
 - Estudiantes y su primer empleo

Contenidos

- Clúster de Investigación Aplicada

- **Proyectos HPC**

- Clúster Hadoop para tecnologías de BI
- Una nube privada para la Administración Pública
- Laboratorios virtuales - *Education as a Service*
- Conclusiones
- Invitación a ERTIC 2013

Proyectos HPC

Actualmente en la FPUNA, 3 focos principales de proyectos:

- Tesis de grado y postgrado
- Investigadores y docentes a tiempo completo
- Convenios de Investigación para el sector público/privado

Proyectos HPC

DSc. Christian Schaerer (chris.schaerer@gmail.com)

- Computación de alto desempeño en simulación y computación paralela.
 - Meteorología, Optimización
- Investigaciones iniciales en GPUs.
 - Métodos numéricos

Proyectos HPC

Proyectos importantes y recientes:

- *Investigación sobre tecnologías para Business Intelligence*
 - Clúster Hadoop para BI
 - Enfoque en la problemática de Big Data
 - Productos Open Source y Comerciales, Appliances
- *Tesis de grado*
 - Una nube privada para la Administración pública
 - Laboratorios virtuales – *Education as a Service*

Proyectos HPC

- Clúster montado

- Server
- 45 Nodos
(*commodity servers*)

Nodos del clúster

- CPU:** Core I5 2400 @ 3.1 GHz 4 núcleos
- Memoria:** 4 GB.
- HD:** 1 TB, Sata III 7200rpm
- Sistema Operativo:** Ubuntu 11.10 x64

Contenidos

- Clúster de Investigación Aplicada
- Proyectos HPC

- **Clúster Hadoop para tecnologías de BI**

- Una nube privada para la Administración Pública
- Laboratorios virtuales - *Education as a Service*
- Conclusiones
- Invitación a ERTIC 2013

Clúster Hadoop para tecnologías de BI

- Proyecto para el sector privado
- *Data Warehouse* y su creciente uso para *Business Intelligence*
- La problemática de “Big Data”
- Necesidad de nuevas tecnologías para su tratamiento
 - Base de datos columnares
 - Clúster
 - Cloud Computing

Hadoop

- Framework para almacenamiento y procesamiento de grandes cantidades de datos sobre un clúster de computadoras
- Conjunto de proyectos de Apache (*Stack Hadoop*)
 - *Hadoop Distributed File System (HDFS)*
 - MapReduce
 - HBase
 - Hive
 - Otros

Hadoop (II)

Ventajas

- Altamente tolerable a fallos
- Linealmente escalable
- Manejo de grandes cantidades de datos
- Software Libre
- Funciona sobre *commodity servers*

Desventajas

- Dificultad de escribir *Jobs MapReduce*
- Ubicación de los bloques de datos no es conocida por las aplicaciones clientes
- No posee un lenguaje declarativo de consultas
- Complejidad administrativa del clúster

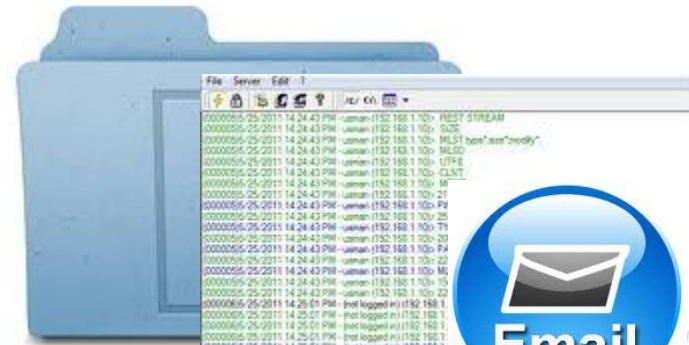
Hadoop (III)

- Hbase
 - Base de Datos Distribuida orientada a columnas.
 - HDFS + MapReduce.
 - No relacional y no soporta SQL.
- Hive
 - Proyecto de DW que funciona sobre Hadoop.
 - Consultas a través de HiveQL.
 - Lenguaje parecido a SQL.
 - Las consultas se traducen a tareas MapReduce

Datos no estructurados



No se almacenan en una base de datos relacional



Se almacenan en documentos de múltiples fuentes

Género	Presión Arterial	Peso	Código de Enfermedad
M	175	65	3
F	141	75	1
...
F	160	59	2



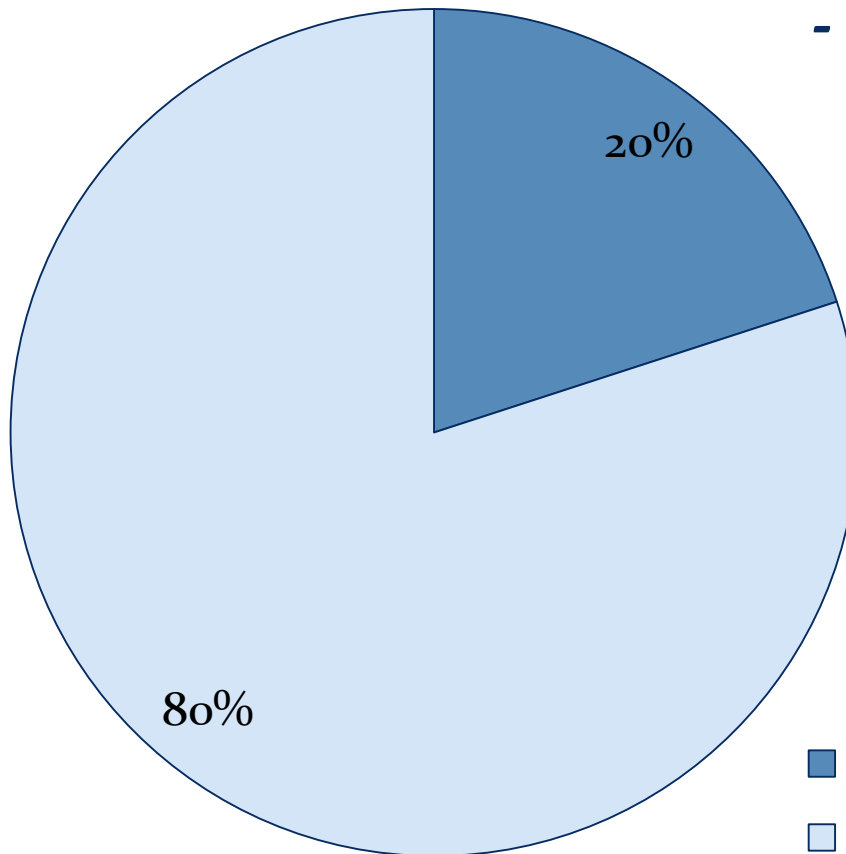
No tienen un formato o estructura predefinida

No se encuentran ordenados

Difícil de categorizar o clasificar para su uso en BI

La regla del 80%

- *Seton Healthcare Family*
- *Merrill Lynch*



- Datos Estructurados
- Datos No Estructurados

Estructurar lo no estructurado

- Identificar patrones.
- Convertir a un formato con el cual se pueda trabajar.
- Agregar estructura, significado:
 - ❑ Extraer información.
 - ❑ Anotaciones o etiquetas.
 - ❑ Análisis de contenido.

Hadoop Streaming

log procesado

1340749576.209|10.13.6.196|http://www.educa.una.py/politecnica/lib/editor/tinymce/tiny_mce/3.4.6/themes/advanced/skins/o2k7/ui_silver.css|jorgeramirez

1340884765.947|10.13.0.37|http://www.educa.una.py/politecnica/lib/editor/tinymce/tiny_mce/3.4.6/theme/s/advanced/skins/o2k7/ui_silver.css|juan

1340749576.209 22 10.13.6.196
TCP_MISS/200 1434 GET
http://www.educa.una.py/politecnica/lib/editor/tinymce/tiny_mce/3.4.6/themes/advanced/skins/o2k7/ui_silver.css jorgeramirez
ROUNDROBIN_PARENT/10.1.5.151
text/css

1340884765.947 172 10.13.0.37
TCP_MISS/200 1435 GET
http://www.educa.una.py/politecnica/lib/editor/tinymce/tiny_mce/3.4.6/themes/advanced/skins/o2k7/ui_silver.css juan
ROUNDROBIN_PARENT/10.1.5.151
text/css

access.log

ETL

almacenamiento

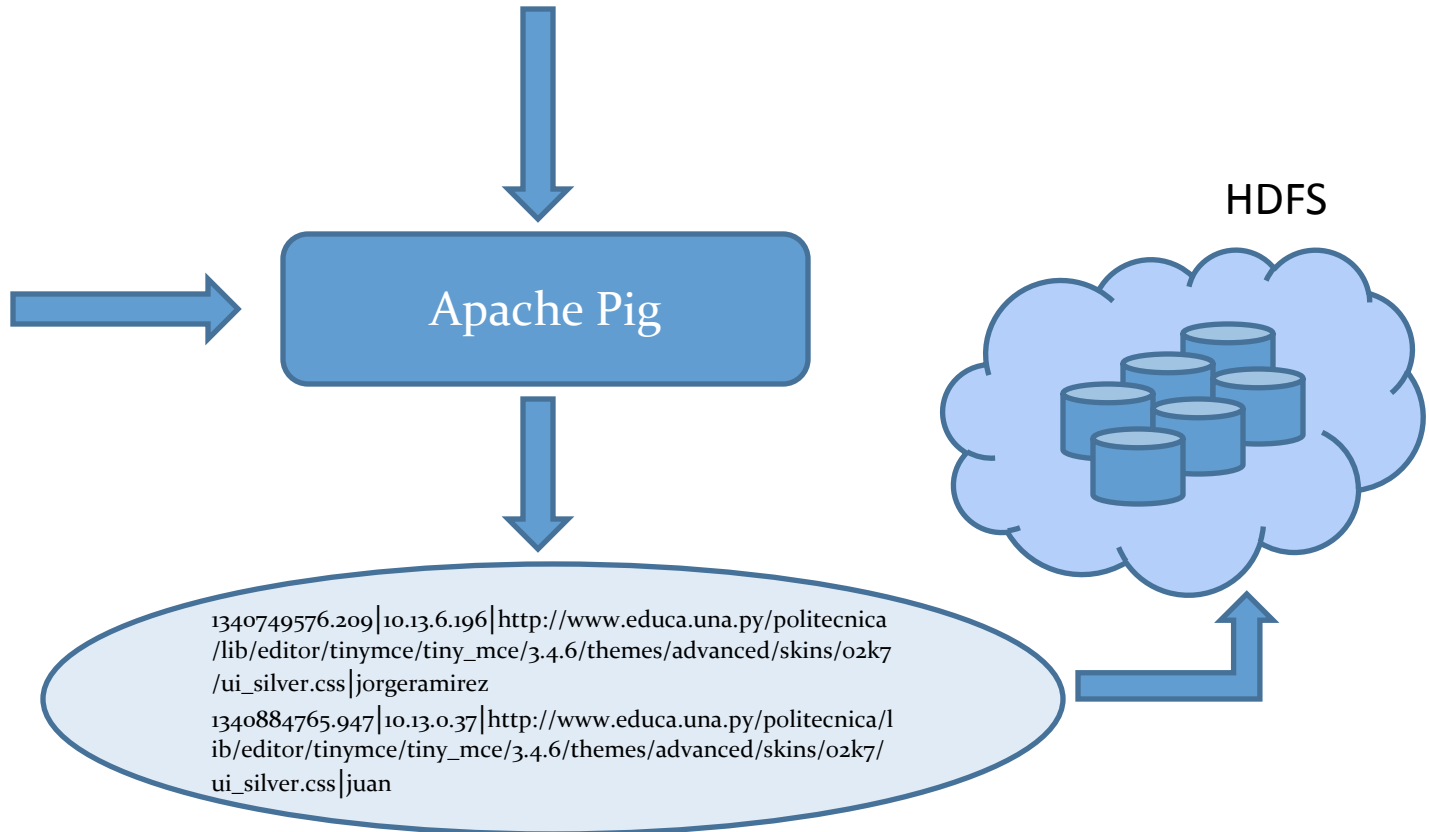
Apache Pig. Procesamiento de log de acceso

Pig Latin

```
A = LOAD '$HDFS_INPUT' USING TextLoader() as line:chararray;  
B = FOREACH A GENERATE FLATTEN(REGEX_EXTRACT_ALL(line,  
'$PATTERN'));  
C = FILTER B BY $0 is not null;  
STORE C INTO '$HDFS_OUTPUT' USING PigStorage('$SEP');
```

```
1340749576.209 22 10.13.6.196  
TCP_MISS/200 1434 GET  
http://www.educa.una.py/politec  
nica/lib/editor/tinymce/tiny_mc  
e/3.4.6/themes/advanced/skins/o  
2k7/ui_silver.css jorgeramirez  
ROUNDROBIN_PARENT/10.1.5.151  
text/css  
  
1340884765.947 172 10.13.0.37  
TCP_MISS/200 1435 GET  
http://www.educa.una.py/politec  
nica/lib/editor/tinymce/tiny_mc  
e/3.4.6/themes/advanced/skins/o  
2k7/ui_silver.css juan  
ROUNDROBIN_PARENT/10.1.5.151  
text/css
```

access.log



Contenidos

- Clúster de Investigación Aplicada
- Proyectos HPC
- Clúster Hadoop para tecnologías de BI

- **Una nube privada para la Administración Pública**

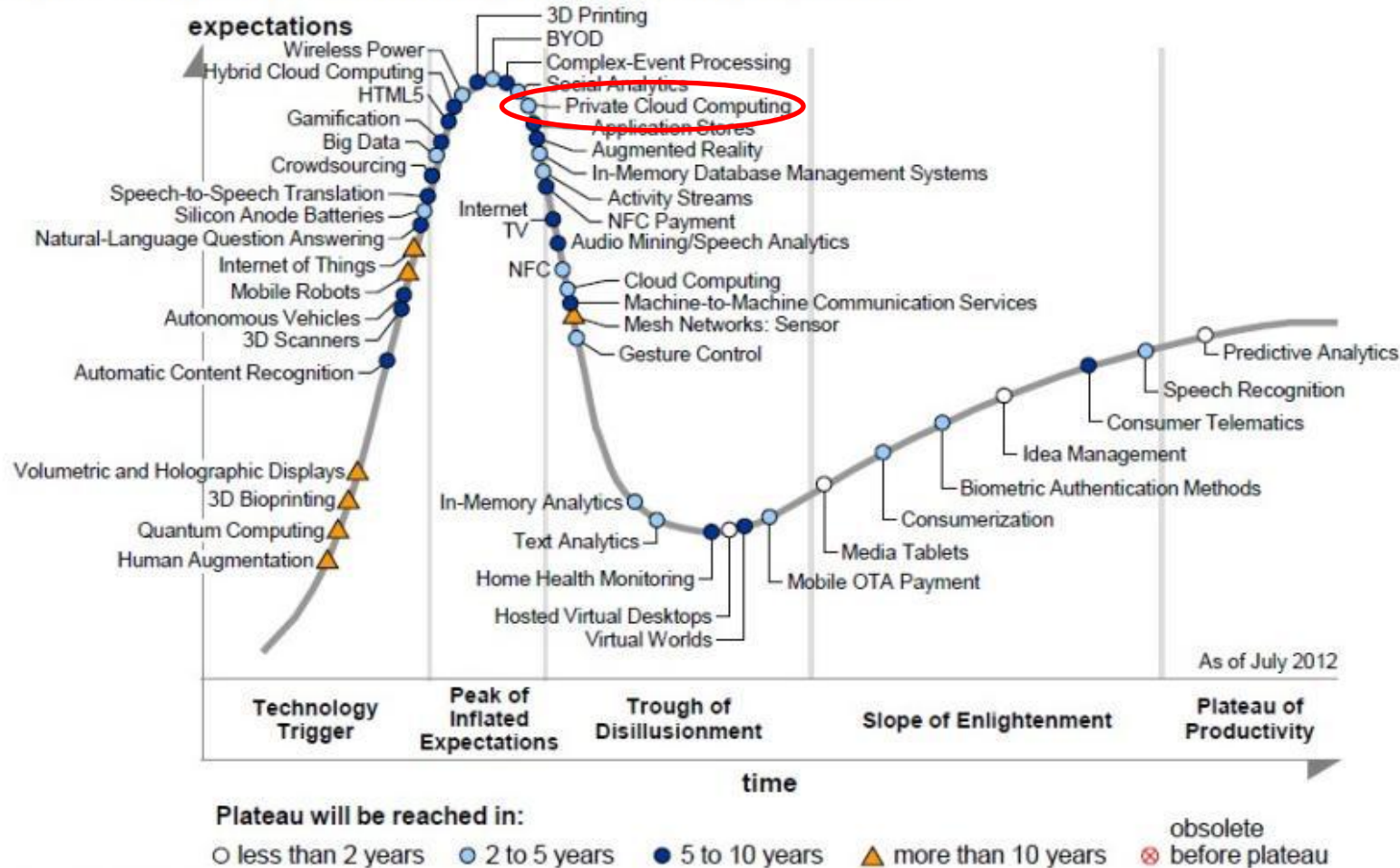
- Laboratorios virtuales - *Education as a Service*
- Conclusiones
- Invitación a ERTIC 2013

Una nube privada para la Administración Pública

- Uso más eficiente de los recursos públicos
- Maximizar el beneficio de las inversiones en TI para los ciudadanos.
- *Open Government*
- Países como Japón, Australia, EEUU, Singapur, Corea, Reino Unido y Dinamarca tienen mayor experiencia en la nube computacional.
- Paraguay a través de la SETICs propone montar su propia nube (privada) para el Gobierno.

La Nube Privada

Figure 1. Hype Cycle for Emerging Technologies, 2012



Source: Gartner (August 2012)

Una nube privada para la Administración Pública

Objetivos

1. Realizar un estudio de los middlewares más importantes de la computación en cloud
2. Diseñar, implementar y probar una cloud, en la Facultad Politécnica de la UNA.
3. Demostrar los beneficios, inconvenientes y limitaciones que representaría para la Administración Pública la adopción de una cloud.
4. Elaborar recomendaciones técnicas para la implementación de la nube en la AAPP del Estado Paraguayo

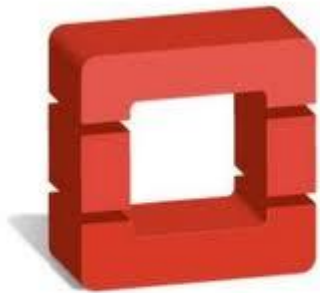
Una nube privada para la Administración Pública

Criterios de evaluación, categorías:

- Características funcionales: Identifica las características funcionales esenciales basado en modelo propuesto por Forrester (2011).
- Seguridad computacional: Identifica las características relacionadas a la seguridad computacional basados en las características mencionadas en Amazon.com, Inc. (2011), IBM Corporation (2011), y el Fraunhofer Research Institute AISEC (2011)
- Calidad y madurez de software libre: Identifica los riesgos desde el punto de vista del usuario y desde el punto de vista del proveedor del servicio respectivamente, basados en los criterios propuestos por QSOS (Atos Origin, 2006).

Una nube privada para la Administración Pública

Middleware evaluados, atendiendo FOSS:



openstack™

cloudstack
open source cloud computing

OpenNebula

EUCALYPTUS



Contenidos

- Clúster de Investigación Aplicada
 - Proyectos HPC
 - Clúster Hadoop para tecnologías de BI
 - Una nube privada para la Administración Pública
- **Laboratorios virtuales - *Education as a Service***
- Conclusiones
 - Invitación a ERTIC 2013

Laboratorios virtuales – *Education as a Service*

*“convertir la educación en **accesible** para todos los miembros de una sociedad, comparable con algo tan básico como el servicio de agua potable, al cual todos puedan tener **acceso consistente y eficiente**”*

El enfoque de **servicio** sugiere
que podría estar en la nube



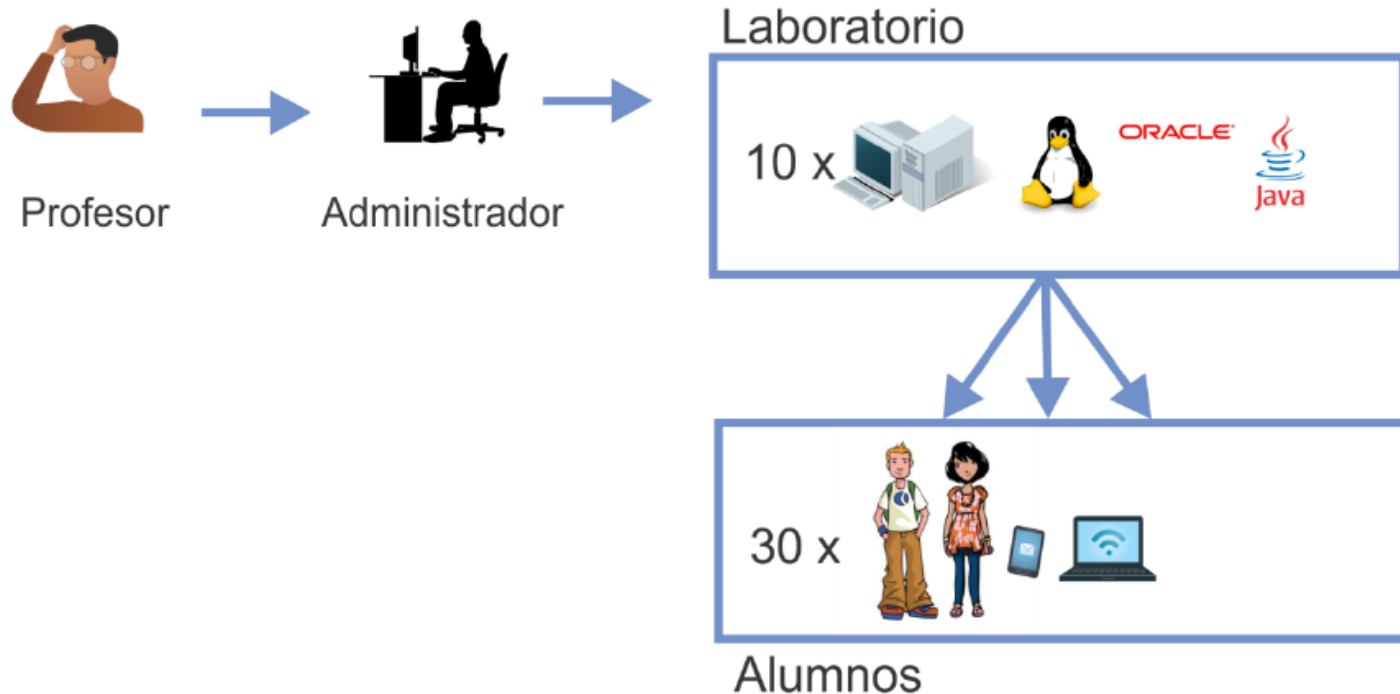
Laboratorios virtuales – *Education as a Service*

Objetivos del trabajo

- Ofrecer la utilización de laboratorios computacionales como un servicio.
- Utilizar arquitectura basada en cloud computing.
- Brindar versatilidad de tecnologías (software) a los estudiantes.
- Brindar alta disponibilidad de herramientas de aprendizaje a los estudiantes.

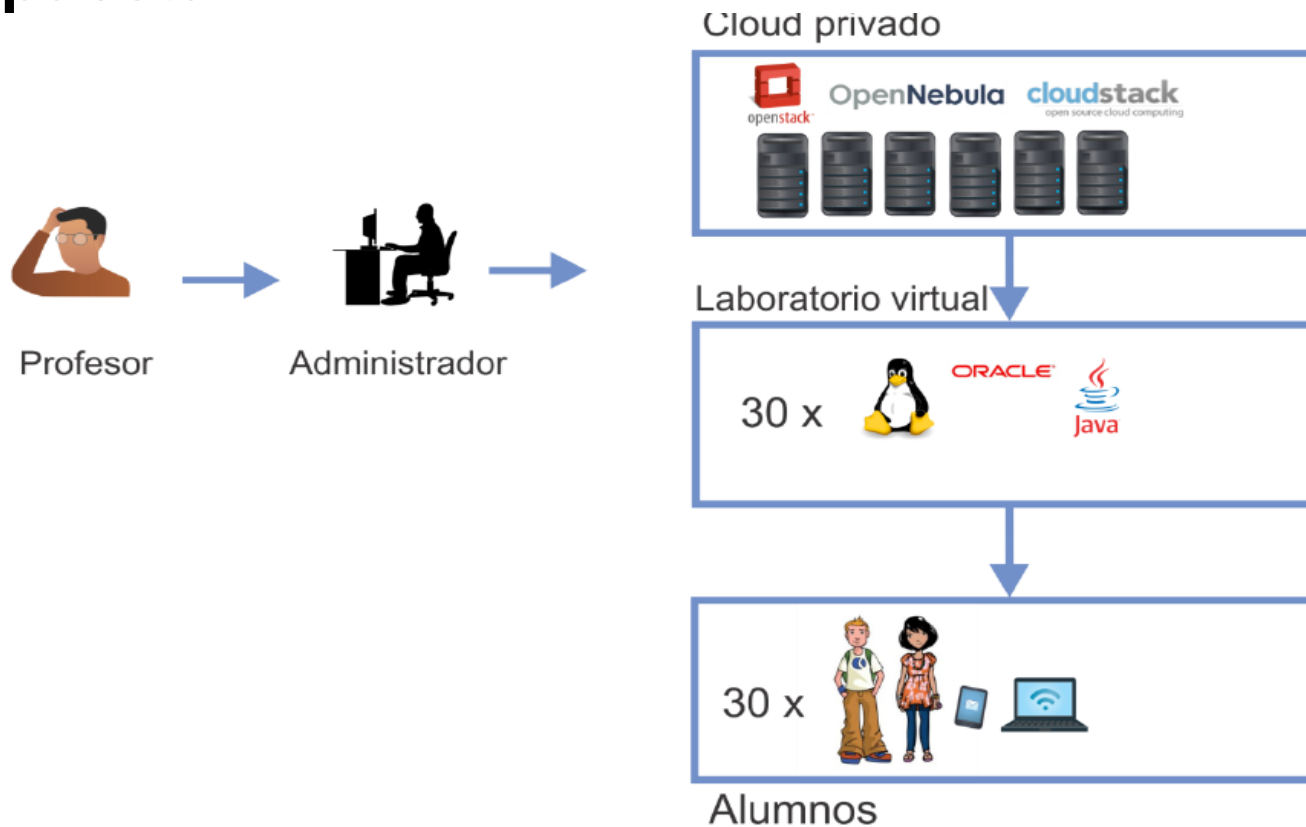
Laboratorios virtuales – *Education as a Service*

Tradicionalmente



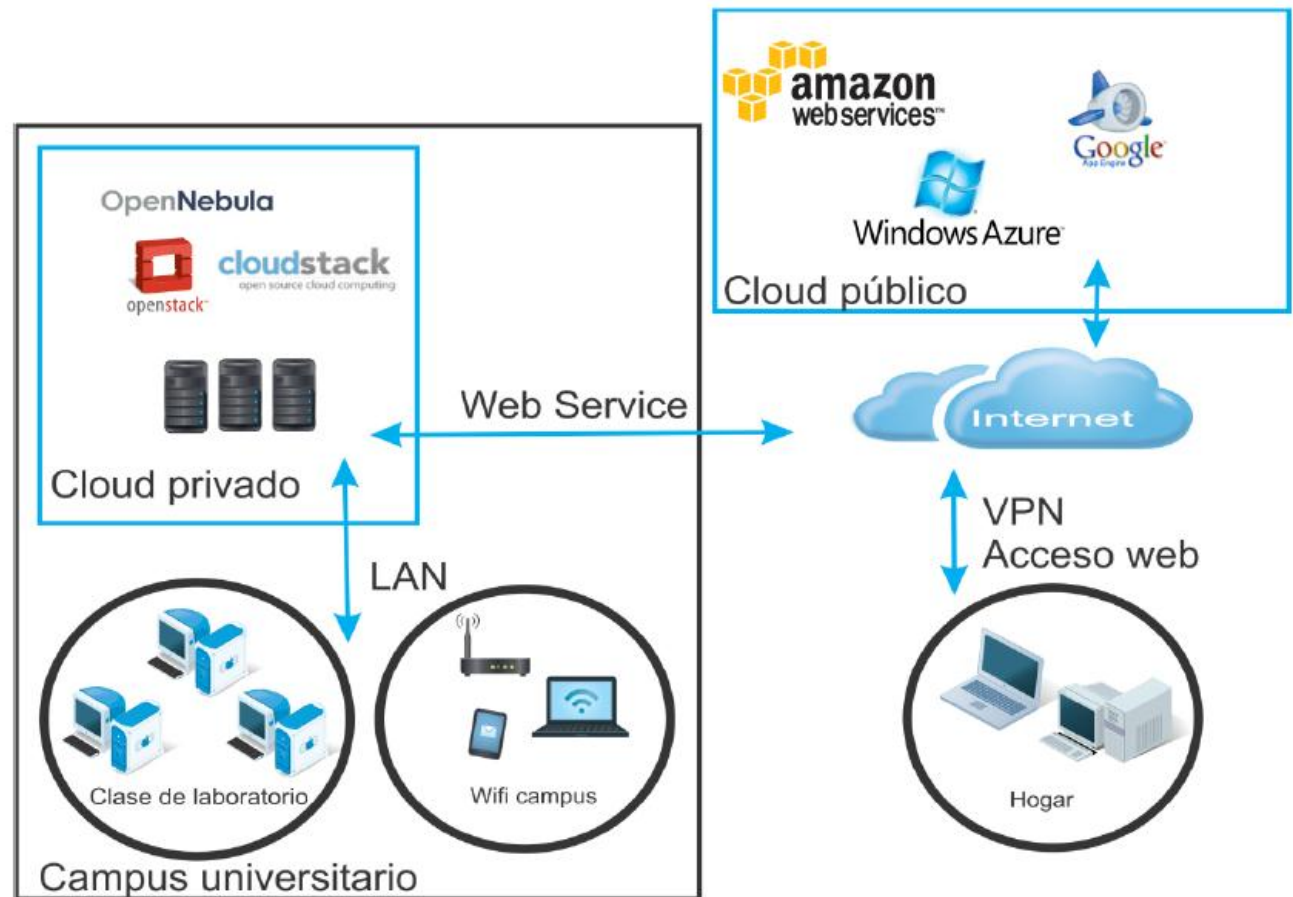
Laboratorios virtuales – *Education as a Service*

Propuesta



Laboratorios virtuales – *Education as a Service*

Arquitectura



Contenidos

- Clúster de Investigación Aplicada
- Proyectos HPC
- Clúster Hadoop para tecnologías de BI
- Una nube privada para la Administración Pública
- Laboratorios virtuales - *Education as a Service*

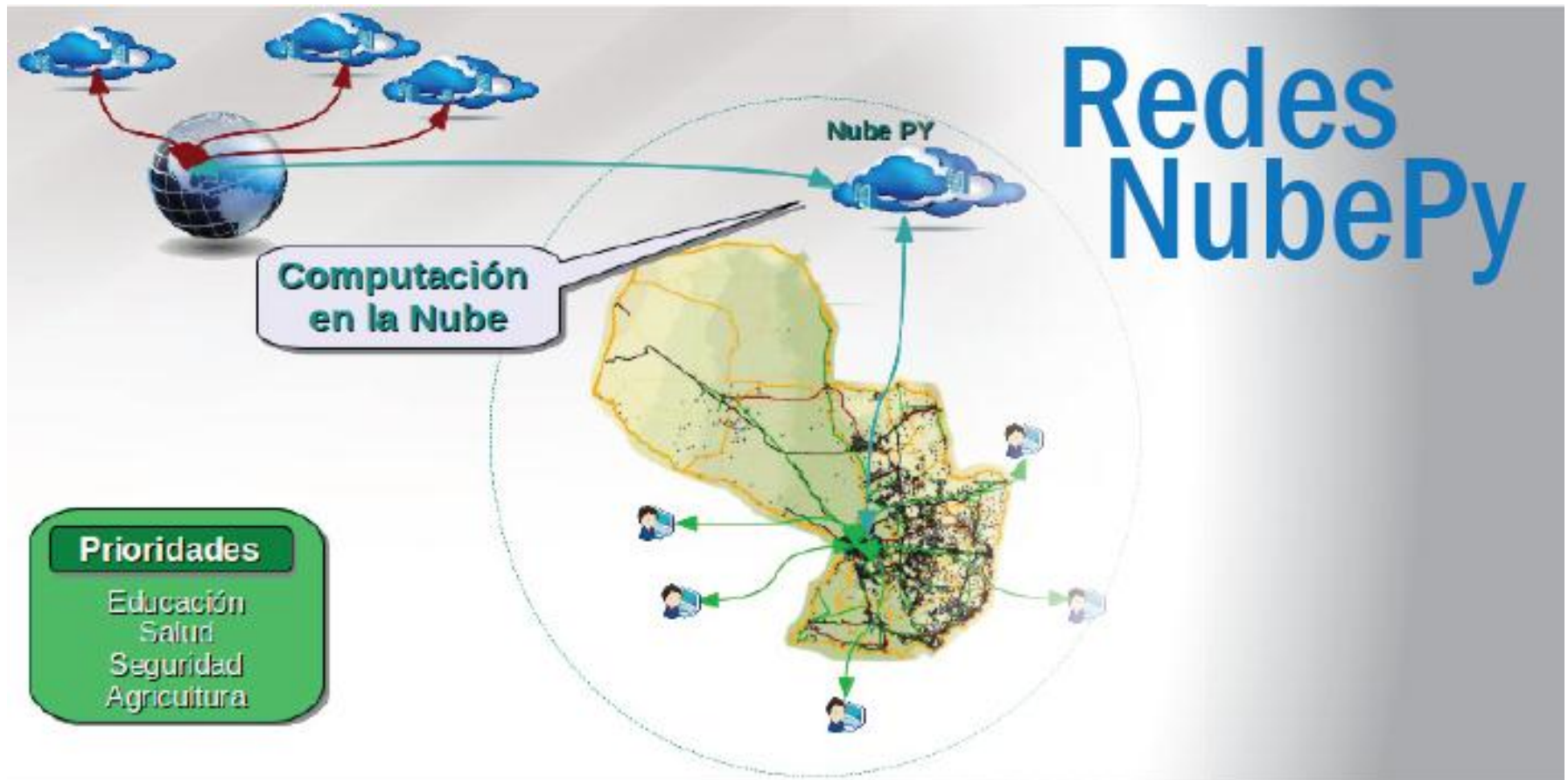
- **Conclusiones**

- Invitación a ERTIC 2013

Conclusiones

- Big Data
 - Procesamiento de grandes volúmenes de datos
 - Clúster, grid & cloud computing
 - Hadoop alternativa software libre
 - Cloudstack, OpenStack alternativas de software libre

Conclusiones



Conclusiones

- *Education as a Service*
 - Acceso para todos
 - Facilitar la gestión
 - Optimización en el uso de los recursos
- Aplicación de Cloud Computing a la Educación

Contenidos

- Clúster de Investigación Aplicada
- Proyectos HPC
- Clúster Hadoop para tecnologías de BI
- Una nube privada para la Administración Pública
- Laboratorios virtuales - *Education as a Service*
- Conclusiones

- **Invitación a ERTIC 2013**



Invitación a ERTIC 2013

- Organizada por la Carrera de Ingeniería Informática de la FP-UNA.
- Se ofrecen cursos intensivos que incluyen una evaluación final.
- Normalmente se realiza en Setiembre del año en curso

Invitación a ERTIC 2013

Cursos ERTIC 2012:

- Recuperación de Información
- Gestión de procesos de negocios usando SOA
- Algoritmos Espectrales y sus aplicaciones
- Web Security
- Técnicas de computación y escalabilidad para la Web
- Introducción al gobierno de las tecnologías de la información: las tareas de un CIO

Invitación a ERTIC 2013

- Invitación para Profesores de Cursos o Conferencistas

Incluye pasajes, hospedaje y viático

- Cursos intensivos, Set/2013:

Lunes a viernes, por la mañana o tarde

- Contacto principal:

Prof. Ing. María Elena García (mgarcia@pol.una.py)