

MINERA DE DATOS

PARA LOGS DE ALMA COMMON SOFTWARE

Javier Fuentes Muoz

j.fuentes06@ufromail.cl

Universidad de La Frontera

28 de enero de 2016



Acerca de ALMA

Colaboración ALMA-UFRO



Acerca de ALMA



...explorando nuestros orígenes cósmicos...

Problemtica general

- Software distribuido
- Altamente complejo
- Grandes volúmenes de logs del sistema
- Análisis de los datos

DRAFT

Objetivo General

- Aplicar metodologas CRISP para generar un modelo de anlisis de los LOGS de ACS.

DRAFT

Objetivos Específicos

- Identificar un problema relevante para ALMA y que pueda ser abordado analizando los logs del sistema ACS.
- Diseñar una metodología basada en CRISP para abordar la problemática identificada.
- Crear un prototipo de pre-procesamiento del log de ACS.

DRAFT

- Infraestructura distribuida
- CORBA
- Sistema de logging

DRAFT

Infraestructura distribuida

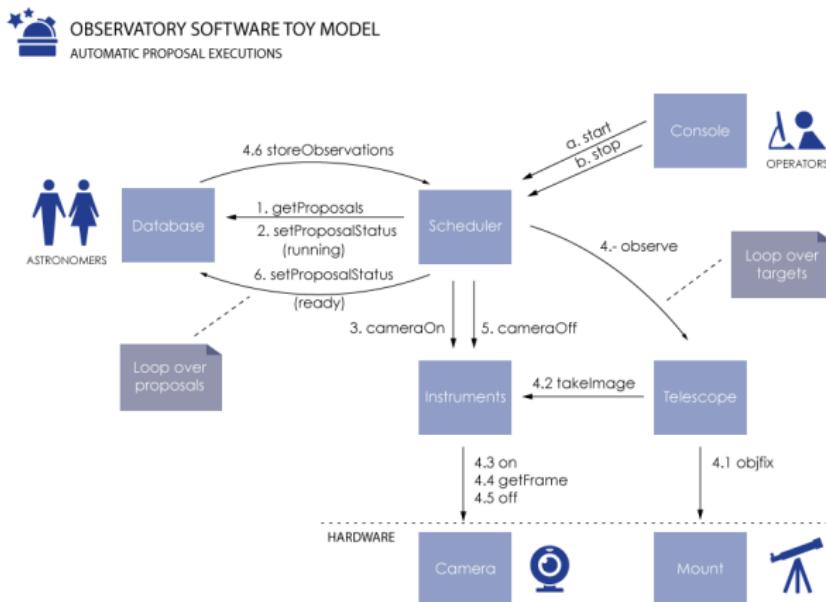


Figura: ALMA-ACS Toy Model

CORBA

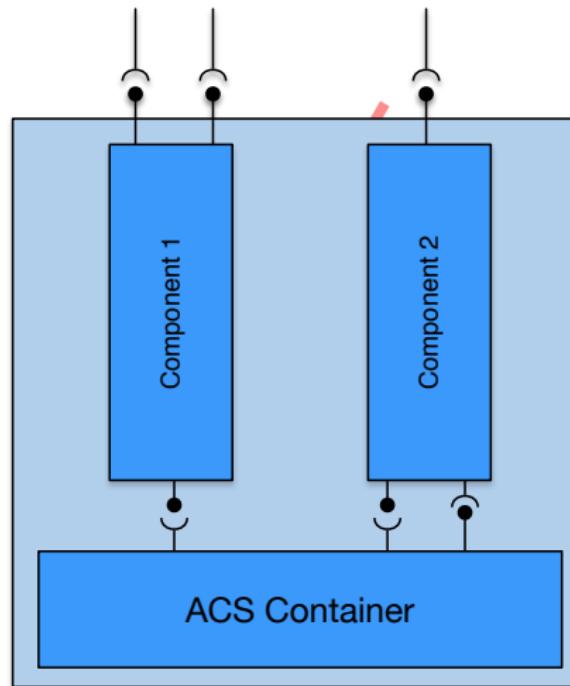


Figura: ACS Container/Component Model

Logging System

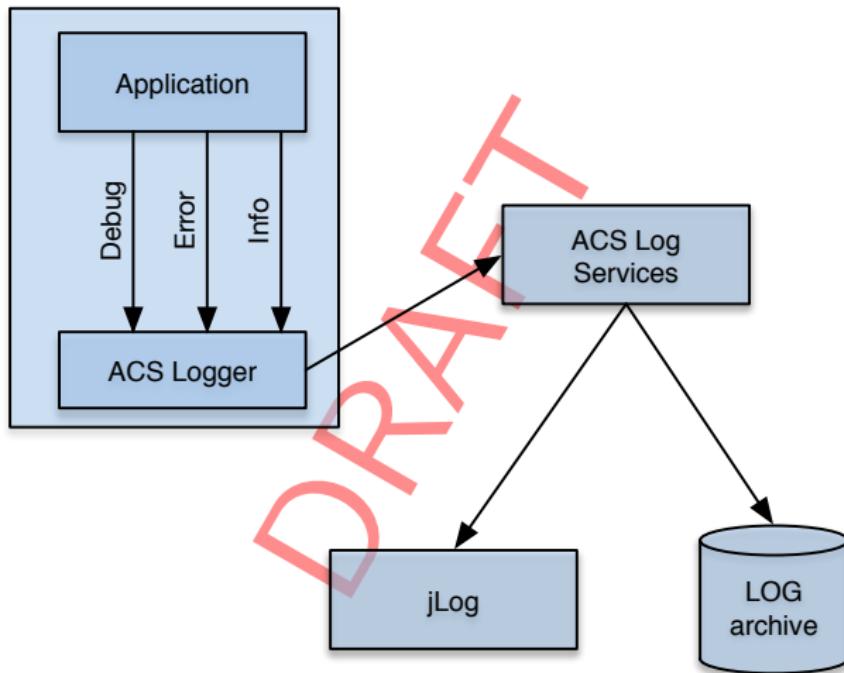


Figura: ACS Logging system

Listing 1: Ejemplo un evento XML de log

```
<Debug TimeStamp="2002-10-7T13:44:16.530"  
^^IHost="te1.hq.eso.org"  
^^IProcess="baciTestServer"  
^^IThread="main"  
^^IContext=""  
^^IFile="baciTestClassImpl.cpp"  
^^ILine="205"  
^^IRoutine="BaciTestClass::~BaciTestClass">  
    Great debug message!  
</Debug>
```

Pre-procesamiento

- Tratamiento de texto para eventos
 - Separar atributos de cada evento.
 - Remover nmeros en los mensajes.
 - Eliminar atributos vacos.
 - Transformar el tiempo a formato UNIX
 - *Stemming* de palabras, (reduccin a la raiz de las palabras)
 - Eliminacin de conectores en mensajes.
- Clustering
 - Distancia de Levenshtein
 - Kmeans - IDF
 - Locality-sensitive hashing algorithms

- SVM

D D

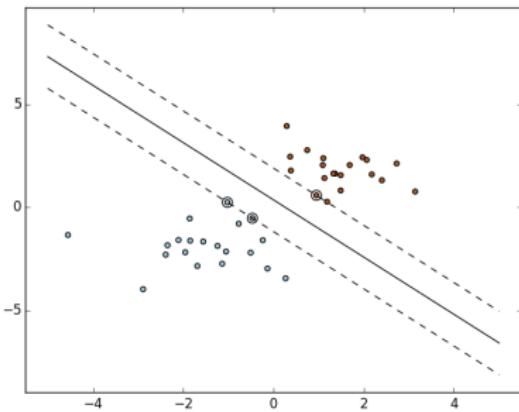


Figura: SVM: Margen de separación

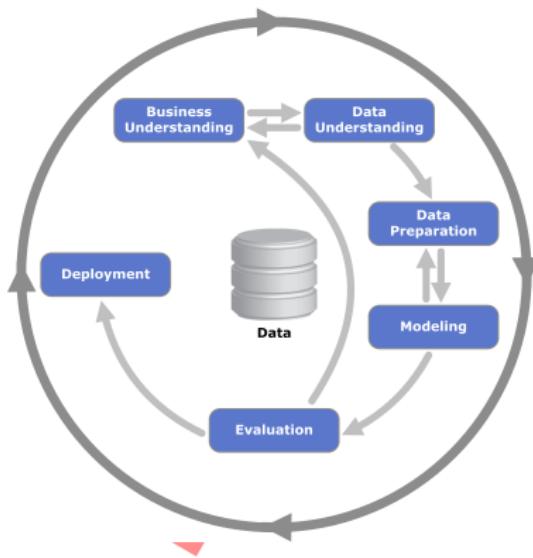
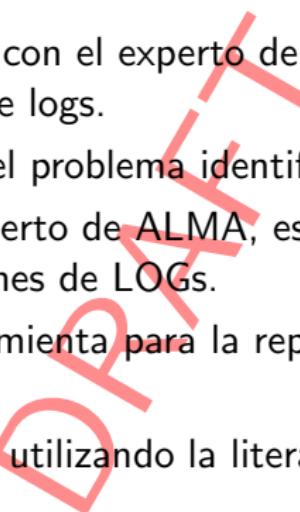


Figura: Proceso CRISP

Programa

- 
- ① Identificar en conjunto con el experto de ALMA una problemtica contingente al anlisis de logs.
 - ② Definir los requisitos del problema identificado.
 - ③ En conjunto con el experto de ALMA, establecer caractersticas relevantes en los patrones de LOGs.
 - ④ Implementar una herramienta para la representacin visual de las secuencias de log.
 - ⑤ Abordar la problemtica utilizando la literatura relacionada con el tema de anlisis de logs.

Programa

- ⑥ Identificar una técnica adecuada para enfrentar la problemática.
- ⑦ Desarrollar un proceso reducido de **KDD**, sin iteraciones, basándose en la técnica previamente seleccionada.
 - ① Selección de un set de datos y operaciones de pre-procesamiento sobre los mismos.
 - ② Extracción de características de las secuencias.
 - ③ Procesamiento de los datos.
 - ④ Validación de los resultados.
- ⑧ Elaboración del proyecto.

Carta Gantt

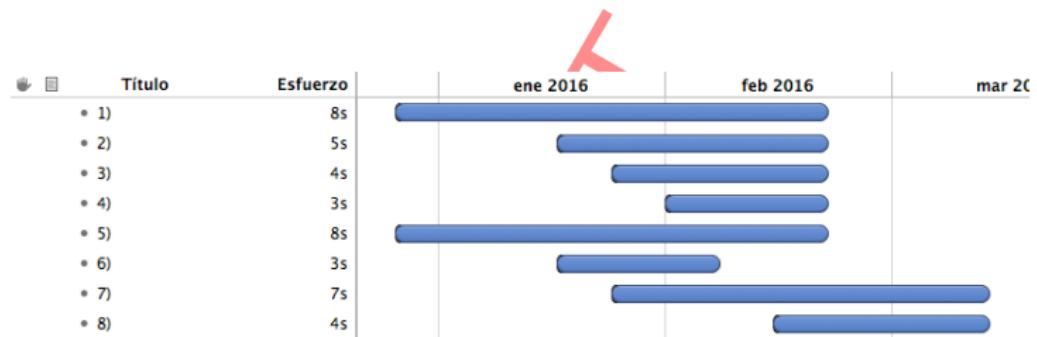


Figura: Carta Gantt

MINERA DE DATOS PARA LOGS DE ALMA COMMON SOFTWARE

Javier Fuentes Muoz

j.fuentes06@ufromail.cl

Universidad de La Frontera

28 de enero de 2016

