



Progetto LTDP di CDF @ INFN-CNAF

- Stato –

Silvia Amerio, Enrico Fattibene, Michele Pezzi

02 novembre 2016

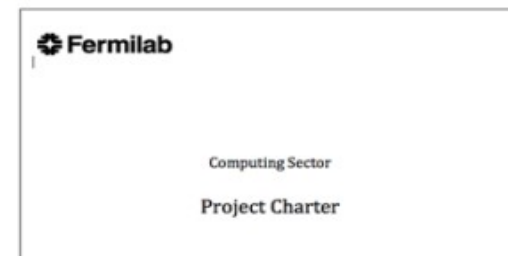
Il progetto DP di FNAL

Il progetto INFN **si affianca al progetto sviluppato a FNAL**, ma si è sempre mantenuto **indipendente da esso**, sia per quanto riguarda i fondi che per il personale coinvolto.

In questa slide gli obiettivi del progetto di FNAL:

*Obiettivi
(dal
Project
charter)*

- Goal: maintain **full analysis capability** (level 4) through **2020** (end of support for Scientific Linux [SL]6)
 - Seek common solutions between experiments where possible
 - Efficient division of effort between experiments and CS
 - Experiments primarily responsible for code, documentation
 - CS primarily responsible for infrastructure (e.g. job submission, databases)
- Does **not** cover (as of now)
 - Accessibility of data by non-collaborators
 - Analysis capability beyond 2020



Data preservation means different things to different people. The approach for this project will be to:

- curate the data from CDF and D0 Run II (both simulated and actual) by continuously migrating the data to modern storage media,
- maintain the full infrastructure ability to generate new Monte Carlo samples, simulate them, reconstruct them, and process them, and
- maintain the ability to perform physics analysis on both the simulated and actual data.

With this, a complete physics analysis can be carried out on Run II data well into the future.

Stato del progetto a FNAL

Il progetto di DP di FNAL è ormai concluso. Qui i punti fondamentali:

Dati

- Migrazione su nuova tecnologia di tape (T10Kd)

Software

- Legacy release basata su SL6 ufficiale
- Distribuita via CVMFS

Sottomissione dei job

- Uso opportunistico delle risorse di FNAL
- Migrazione di CDF al nuovo sistema di sottomissione dei job – *jobsub*, utilizzato dagli altri esperimenti di FNAL

Documentazione

- Riorganizzate le pagine web di CDF
- Migrazione di tutte le note di CDF (pubbliche e interne) su Inspire

Il progetto INFN-CNAF

Obiettivo: preservare una copia completa dei dati e dei campioni MC di CDF (ntuple e dati raw, circa 4 PB) + tutti i servizi per l'accesso e l'analisi dei dati

Motivazione (dal documento del progetto presentato ai referee e approvato in commissione 1)

INFN involvement in long term CDF data preservation is important for different reasons:

- 1) INFN strongly contributed to the success of CDF; data preservation is the last phase of the experiment and INFN should have a role in it.*
- 2) Access to data many years from now, beyond CDF collaboration, will be under INFN control.*
- 3) A mirror archival in Europe is a necessary safety measure.*
- 4) Direct participation with a real case to the problem of data preservation, which is of great interest → CDF preservation system at CNAF can serve as a prototype for future experiments now supported by INFN.*
- 5) Opportunity for CNAF to take a significant role in the long term preservation of data.*

Cronologia

- progetto approvato a **settembre 2012**
- fondi disponibili a **giugno 2013**
- primi tape disponibili a **gennaio 2014**
- copia iniziata a **febbraio 2014**
- copia completata a **febbraio 2015**
- **da febbraio 2015** iniziate le attività di
 - recupero dei file mancanti
 - installazione nuovo sistema di accesso ai dati e nuovo codice CDF
 - analisi dati con il nuovo sistema jobsub (via FNAL)
 - produzione pagina web

Fondi assegnati

2013

Assegnati **89 kEuro**

Ripartizione:

- utilizzo di due tape drive nel periodo 2013-2014: **15 kEuro**
- servers hms: **8 kEuro**
- nastri: **66 kEuro**

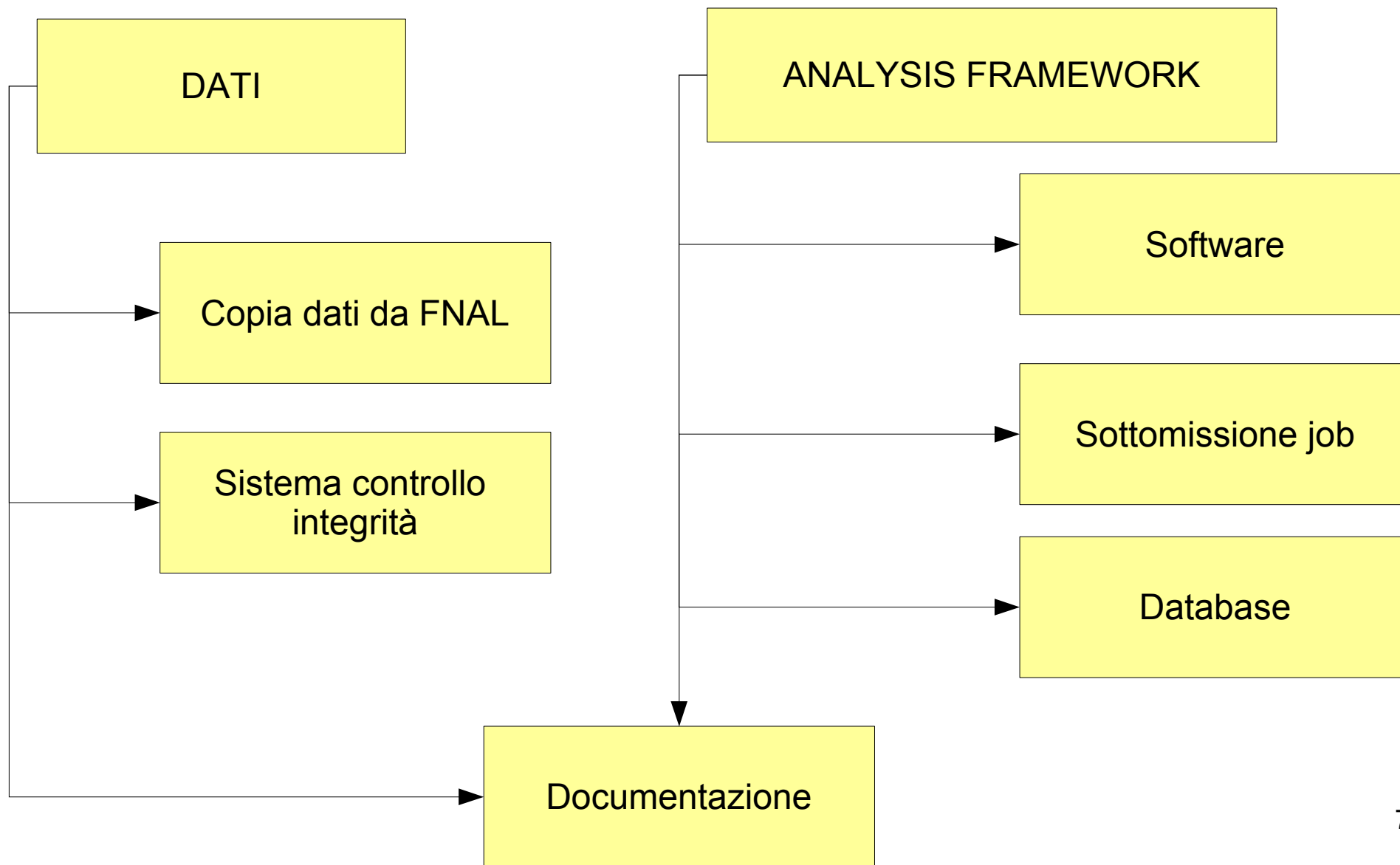
2014

Assegnati **99 kEuro**

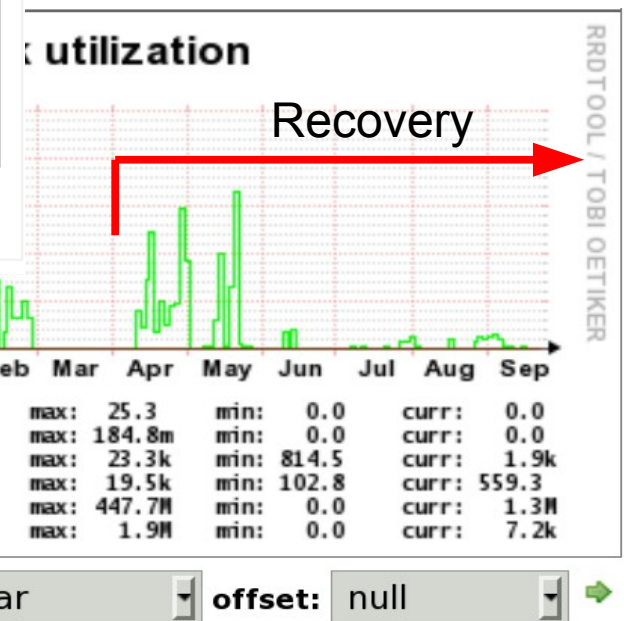
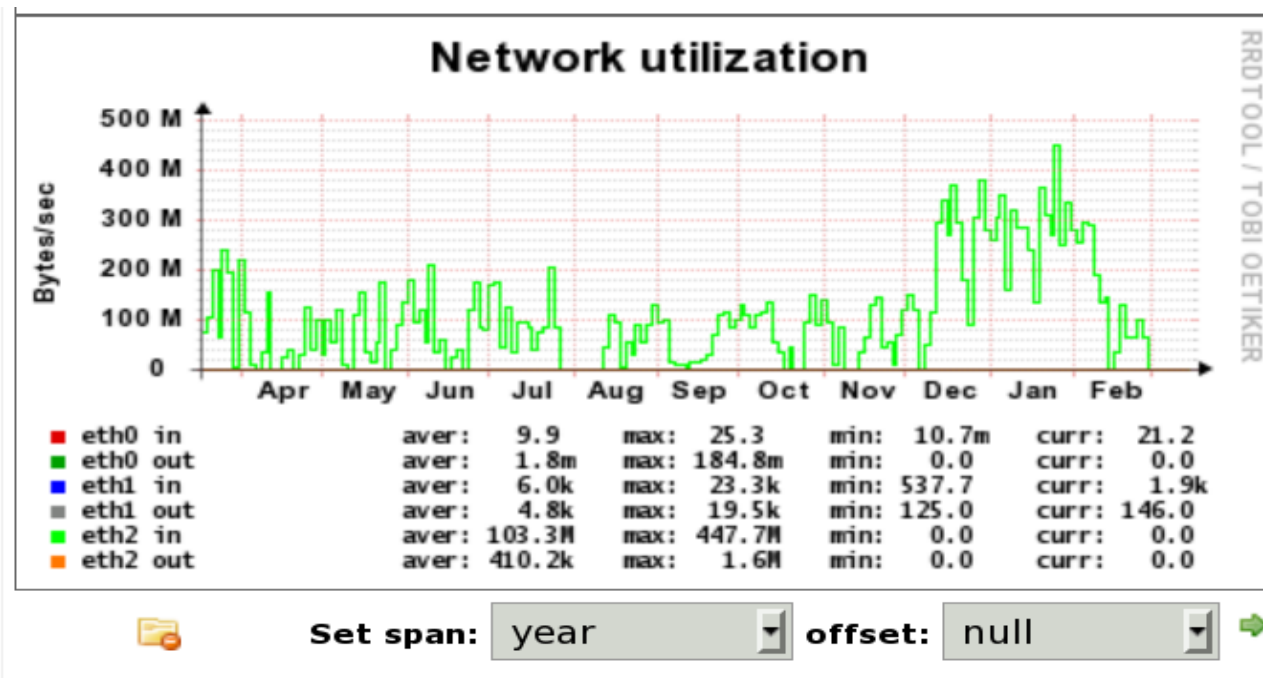
Ripartizione:

- server per il DB: **4 kEuro**
- nastri: **95 kEuro**

Schema progetto



Copia dei dati



Copia completata a Febbraio 2015.

Da aprile 2015 iniziato il recovery delle copie fallite e il check finale di tutti i file. (ad oggi mancano ancora 50k file- circa 50TB)

Attualmente 4,3 PB su tape.

Controllo integrità dei dati

Script per il controllo periodico dei dati

E' in corso il controllo finale di tutti i dataset. Per ogni dataset lo script di controllo:

- controlla se tutti i file sono stati copiati e sono su nastro al CNAF
- se manca qualche file, automaticamente avvia la copia da FNAL
- per i file presenti, controlla se il CRC è presente tra gli attributi estesi del file e se coincide con il valore archiviato nel DB a FNAL
- se il valore è assente, rilegge il file da tape e aggiunge il valore CRC agli attributi del file
- se il valore è diverso da FNAL, il file viene nuovamente copiato.

La versione dello script per il futuro a lungo termine prevede anche il richiamo dei file da tape.



Accesso ai dati

Nuovo sistema di accesso ai dati (SAM web)

FNAL ha sviluppato una nuova versione del sistema di accesso ai dati (SAMWeb), che usa http per comunicare con il web server del DB. Questa versione – usata anche dagli altri esperimenti di FNAL – sarà supportata nel futuro a lungo termine.

Il nuovo sistema è stato installato al CNAF e testato con successo.

Il codice di CDF

CDF legacy release

La legacy release del codice di CDF si basa su SL6 e a FNAL viene distribuita tramite CVMFS. Al CNAF abbiamo deciso di procedere in due fasi:

- accesso dei nodi SL6 del CNAF al server CVMFS di FNAL via squid (fase completata)
- Installazione al CNAF di una replica del server CVMFS di FNAL, in modo da avere in loco una copia del codice (non ancora effettuata)

La legacy release è stata rilasciata ufficialmente da FNAL a febbraio 2015. E' stata testata con successo sui nodi SL6 del CNAF.

Sottomissione job

JobSub

CDF è passato ad un nuovo sistema di sottomissione dei job, usato anche da altri esperimenti a FNAL.

Il passaggio al nuovo sistema avviene in due fasi:

- Configurazione della macchina di sottomissione di FNAL per inviare job al CNAF (fase completata e testata con successo);
- Installazione e configurazione del sistema jobsub al CNAF (non ancora effettuata).



Database

Non abbiamo ancora copiato al CNAF i database (online e offline e dati) di CDF.
I database si basano su Oracle, al momento CDF (FNAL) non ha intenzione di migrare ad un software libero.
L'ostacolo principale alla copia è quindi per ora il costo della licenza.

Documentazione

Pagina web con la descrizione del progetto e le istruzioni per gli utenti

Long Term Data Preservation CDF@CNAF

Menu :

[Home](#)

[The project](#)

[Bit preservation](#)



Dataset	Count	Warning	Error	Mismatch	Exceedeed	Different	Last check
ahgs1r	9/9	no	no	0	0	0	2015-07-24
ahgs1s	9/9	no	no	0	0	0	2015-07-24
ahgs70	123/123	no	no	0	0	0	2015-07-24
ahgs71	114/114	no	no	0	0	0	2015-07-24
ahgs7a	104/104	no	no	0	0	0	2015-07-24
ahgs7b	106/106	no	no	0	0	0	2015-07-24
ahgs7c	105/105	no	no	0	0	0	2015-07-24
ahgs7d	105/105	no	no	0	0	0	2015-07-24
ahgs7e	199/199	no	no	0	0	0	2015-07-24
ahgs7f	107/107	no	no	0	0	0	2015-07-24
ahgs7g	106/106	no	no	0	0	0	2015-07-24
ahgs7h	108/108	no	no	0	0	0	2015-07-24
ahgs7i	106/106	no	no	0	0	0	2015-07-24
ahgs7j	107/107	no	no	0	0	0	2015-07-24
ahgs7k	107/107	no	no	0	0	0	2015-07-24
ahgs7y	105/105	no	no	0	0	0	2015-07-24
ahgs7z	115/115	no	no	0	0	0	2015-07-24
ahgt10	4/4	no	no	0	0	0	2015-07-24
ahgt11	4/4	no	no	0	0	0	2015-07-24
ahgt12	4/4	no	no	0	0	0	2015-07-24

Pagina web con la descrizione dei dataset al CNAF (creata automaticamente dallo script che controlla i dati)

A breve integreremo anche la descrizione del dataset.

Persone attualmente coinvolte xxx

CDF-Italy

S.Amerio

CNAF

A.Cavalli, L.dell'Agnello, S.Dal Prà, E.Fattibene, M.Pezzi,
P.Ricci

CDF-Fnal (supporto per SAMWeb e Jobsub)

J.Boyd, W.Sakumoto

Fermilab computing sector (supporto per la copia)

D.Litvinsev

Conclusioni

I dati di CDF (~4 PB, dati e MC, formato raw e ntuple) sono stati copiati al CNAF. E' in corso il controllo finale di tutti i dataset per recuperare eventuali file mancanti.

E' in **fase di revisione** il sistema di controllo periodico dell'integrità dei dati.

Il sistema di accesso ai dati e di sottomissione dei job sono stati aggiornati alle versioni (SAMWeb e jobsub) sviluppate a FNAL, che saranno supportate nel futuro a lungo termine.

La legacy release del codice di CDF è accessibile al CNAF tramite CVMFS.

Da completare:

- installazione del server CVMFS e dell'headnode di jobsub
- copia dei DB
- documentazione