INAF-Osservatorio astronomico di Torino

Technical Report nr.155

Gaia
Data Access and Analysis System

Approccio sistematico alla realizzazione di una visualizzazione dei dati funzionale alle attività di valutazione scientifica dei risultati.

Pubblicazione Tecnica n. 155

R. Morbidelli[1] *e E. Pigozzi*[2]

Pino Torinese, 10 ottobre 2011

[1] **INAF – Osservatorio Astronomico di Torino**
[2] ALTEC - **Advanced Logistics Techn. Eng. Center**

Sommario

Antefatto

Questo documento ha lo scopo di fornire un primo approccio, sistematico, nella costruzione di un ambiente costituito da strumenti di visualizzazione dei dati il cui principale uso sia quello di poter, monitorando i processi che avvengono e collezionando i risultati che si rendono disponibili nelle attività che avvengono durante le operazioni di AVU, fornire gli elementi utili al personale scientifico per una valutazione diagnostica e critica di quanto avviene durante il decorso della missione Gaia. Questo con la finalità non solo di visualizzare i processi in intervalli di tempo predefiniti ma anche di analizzare e valutare i risultati che in questo accadimento si producono.

Il documento è indirizzato, per questo motivo, ad una suddivisione di questo processo in cinque casi fondamentali:

1. Visualizzazione dei processi "Run time".
2. Archiviazione e recupero, sistematico, a mezzo di grafici e log files di informazioni contenute nel LDB.
3. Ottenere "On demand" insiemi di informazioni non usualmente messi a disposizione dal processo di popolamento del LDB e poter, di questi, produrre gestioni tali da produrre "off line" una visualizzazione, manipolazione ed analisi dei dati
4. Produzione di tabelle a partire da sequenze storiche di dati presenti non necessariamente nel LDB ma anche in ambiti diversi da quelli connessi al ciclo di processamento portato avanti dalla "Pipe Line"
5. Operazioni non preventivabili di "data Mining" nel LDB e MDB

I differenti casi sono descritti in termini di metodi ed ambienti in cui questi hanno luogo, con la condizione che gli eventi possano essere realizzati, anche in condizioni di criticità, senza impatto critico con l'esecuzione delle procedure concorrenti, e che ove questo avvenga, sia possibile ai fini della valutazione dell'opportunità di procedere disporre di metodi di valutazione sufficientemente affidabili in termini di predizione del carico e dei tempi.

Informazioni di rilievo

I contenuti di questa nota tecnica sono correlati ai contenuti ed obbiettivi del Main WP: GWP-M-CNN-00000

I contenuti di questa nota tecnica sono funzionali alle finalità e funzionamento dei WPs: Tutti i Work Packages di AVU e per l'esperimento GareX

Documenti applicabili

- MM-002 DPCT Development Plan
- JH-001 Main Database Interface Control Document
- RM-002 DPCT Internal Interface Control document

Acronimo	Descrizione
AF	Astrometric Field
AGIS	Astrometric Global Iterative Solution
BAM	Basic-Angle Monitoring (Device)
CCB	Configuration Control Board
CMDB	Configuration Management DataBase
CU	Coordination Unit (in DPAC)
DB	Database
DMS	Document Management System (ESA)
DP	Data Processing
DPAC	Data Processing and Analysis Consortium
DPACE	Data Processing and Analysis Consortium Executive
DPC	Data Processing Centre
DPCE	Data Processing Centre ESAC
DPCT	Data Processing Centre Torino
ESA	European Space Agency
ESAC	European Space Astronomy Centre (VilSpa)
ESTEC	European Space research and TEchnology Centre (ESA)
FL	First Look
FLOP	Floating Point Operation
GPDB	Gaia Parameter Database
GSR	Global Sphere Reconstruction
GTS	Gaia Transfer System
HW	Hardware (also denoted H/W)
IAR	Implementation Acceptance Review
ICD	Interface Control Document
IDL	Interactive Data Language
IICD	Internal Interface Control document
IRD	Interface Requirements Document
LDB	Local Database
LPDB	Local Personal DataBase
MDB	Main DataBase
MOC	Mission Operations Centre
MDB	Main Database
PDB	Personal Database
ScOM	Scientific Operations Manager
SVN	Subversion
WP	Work Package
WPM	Work Package Manager

Considerazioni Preliminari

Esperienza ed uso dimostrano che efficienti ed intuitivi metodi per la presentazione, comprensione e archiviazione dei contenuti e definizioni di un insieme di dati scientifici sono dati dall'adozione sistematica, in ultima analisi, di tre metodi: grafici, immagini e tabelle. La realizzazione di questi, nel caso di Gaia, va condotta tenendo conto dei tempi di processamento dei dati, soggetti ad un incremento in termini dimensionali e di complessità nel tempo, e del destinatario che dovrà farsi carico della interpretazione, valutazione e stima della significatività dei medesimi. Questo fatto implica la possibilità di produrre, in un sistema integrato con i metodi di processamento, in modo standardizzato e non, tutte quelle forme di resoconti orientati a garantire, in forma di riassunto, ovvero estensiva, gli esiti di un processo.
Ne consegue che l'analisi darà luogo a immagini, grafici, tabelle, log files, ecc i cui contenuti saranno, a seconda dei casi, di facile comprensione da parte di chiunque, ove la veste sia sufficientemente compendiata da informazioni per la lettura, ovvero interpretabili esclusivamente da specialisti, auspicabilmente solo nei casi in cui effettivamente i destinatari siano, esclusivamente, questa categoria di utenti.
Prerequisito per poter, comunque, costruire un siffatto sistema è la capacità da parte del medesimo di produrre tre tipologie di azioni:

1. Eseguire una rappresentazione logica e quantitativa di un campione dei dati (es. il grafico).

2. Riempire un file con dati selezionati (ad es. una tabella di un DBMS ovvero un file autoconsistente).

3. Fornire un riassunto delle operazioni compiute (ad es. un report contenente risultati statistici, indici o comunque informazioni compendiate)

La parte Italiana di elaborazione dei dati si ritiene imponga che questi tipi di visualizzazione dei dati divengano disponibili, ben prima del lancio, sia nel contesto del monitoraggio, in un'ottica volta alla verifica del corretto funzionamento operativo dell'infrastruttura, argomento che non è oggetto di questa nota, sia nel contesto di offrire strumenti di: prevalutazione, manipolazione ed analisi dei risultati che vengono a prodursi nel tempo secondo un'ottica d'indagine scientifica. Questo sia per garantire questa opportunità dopo il lancio sia quale ausilio per una comune comprensione, da parte di quanti collaborano alla realizzazione degli strumenti di elaborazione, dei dati stessi.

Per questi ultimi deve essere possibile, sia durante le fasi antecedenti che durante lo svolgimento della missione, con particolare enfasi per le prime fasi della stessa; con vari livelli di incisività; e in coerenza con adeguati livelli di autorità di chi li esegue; interagire con i vari contesti di collocamento dei dati.

Questo ove indispensabile anche nel corso della produzione e dell'archiviazione dei dati stessi. In particolare occorre attribuire il giusto peso al fatto che i dati stessi sono organizzati e quindi principalmente accessibili in virtù degli schemi derivanti da un ben specifico Data Model le cui proprietà sono codificate nell'Internal Interface Control Document del Database (IICD) decontestualizzati tali dati sono sostanzialmente incomprensibili. Quindi i dati sono, in ultima analisi, sempre accessibili avvalendosi di un DataBase Management

System (Oracle DBMS) il cui utilizzo, però, al di fuori dei protocolli previsti per l'infrastruttura, porta ad inevitabili interazioni con i processi di elaborazione.

In questo documento proprio nell'ottica di consentire la riduzione al minimo dei rischi derivanti da queste interazioni viene introdotto il concetto di utilizzo di "Personal Database" locali (local PDB) intesi come ambienti orientati alle attività scientifiche che possano costituire un ottimale compromesso tra l'accesso alle grandi moli di dati che vanno accumulandosi nel "Local DataBase" (LDB) del DPCT con il procedere del processamento in via ordinaria sull'infrastruttura e la possibilità da parte dello scienziato di utilizzare metodi, anche inusuali, di manipolazione dei dati stessi in un ambiente di elaborazione predefinito se non proprietario.

In questo documento tale ambiente viene identificato, per quanto concerne i metodi per sviluppare procedure non standard, principalmente, ma non esclusivamente, nell'utilizzo del software IDL[3].

Al fine di introdurre, inoltre, un'univoca convenzione di identificazione di alcuni dei soggetti della nota, si indica con il termine: "attore scientifico" o più brevemente "attore" una persona con competenze scientifiche (eventualmente non competente di tecniche di DB Management) dell'Osservatorio Astronomico di Torino che abbia motivo di interagire direttamente o indirettamente a vario titolo e per motivazioni diverse con la base dati del DPC-T. Definiamo anche come: MDB, il sito ultimo di collocamento di tutti i dati che via via, nel prosieguo della missione, vanno a costituire l'archivio globale sia dei dati provenienti da ESAC, sia dei risultati dei vari processi di elaborazione. Il riempimento di questa struttura avviene ad intervalli temporali non definibili a priori, ma la somma a fine missione corrisponde, per certo, al totale di tutta l'informazione ricevuta e prodotta. Definiamo LDB il sito di collocazione di dati provenienti nel corso di un ciclo della missione da ESAC e dei risultanti da processamento legati ad uno degli intervalli temporali che caratterizzano le attività del satellite, questa base dati può essere riempita per effetto di eventi che possono avere scale temporali variabili e virtualmente di ordine illimitato: giornaliero, settimanale, mensile, ma che nei fatti non supera il limite teorico dei 6 mesi attribuito, attualmente, ad un ciclo di lungo termine delle procedure AVU.

Da ultimo definiamo LPDB un sito di collocamento dei dati provenienti da ESAC e/o dai processi cui possono accedere gli attori, tale sito può essere unico, ovvero molteplice nel caso ciascun attore riceva un sottoinsieme di dati come effetto dell'output di processi di sua pertinenza ovvero queste informazioni siano il risultato del recupero di dati dal LDB o dal MDB.

---

[3] L'utilizzo di questo tipo di s/w ha, ad esempio, il pregio di fornire ambienti già ampiamente collaudati nell'ambito scientifico Astronomico quali le Librerie della NASA e gli strumenti di elaborazione dati costituiti da: IDL Advanced Math and Stats Module capabilities che includono le classi matematiche e statistiche sintetizzate di seguito:

Mathematical Functions: Linear systems, Eigensystem analysis, Interpolation and approximation, Differential equations, Transforms, Nonlinear equations, Optimization, Matrix/vector operations.
Statistical Functions: Basic statistics, Regression Correlation & covariance, Categorical & discrete data analysis, Nonparametric statistics, Goodness-of-fit and randomness, Time series and forecasting, Multivariate analysis, Survival analysis, Probability distribution functions, Inverses, Random number generation and utilities.

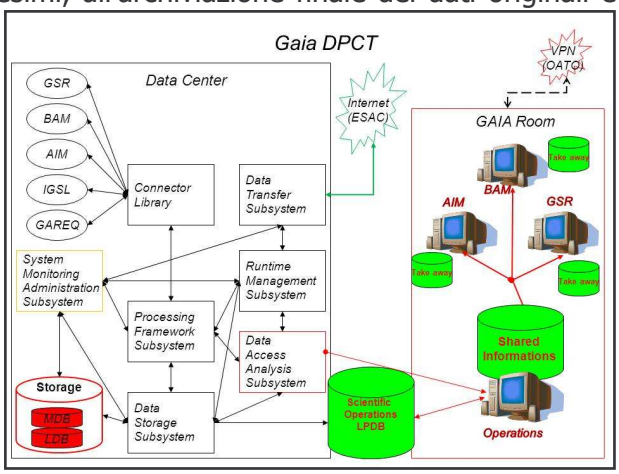
Inoltre la presenza di tools di Data Mining consente un potenziale e persistente link tra quest'ambiente e quanto accantonato nell'Oracle DBMS (versione 11G) adottato per il MDB, il LDB ed ove necessario per la produzione di PDB

Architettura del sistema

Alla luce della necessità di garantire, primariamente, la continuità e l'affidabilità delle attività di data processing del DPC-T sembra irrinunciabile il concetto di isolamento e distinzione dei contesti operativi discriminando tra: ambito di gestione operativa dei dati per il funzionamento della catena di processamento (il data center) e l'ambito di gestione orientata alla valutazione della significatività scientifica dei dati prodotti, sia con mezzi automatici o semiautomatici che manualmente mediante procedure "on demand" (Gaia room o risorse di calcolo individuali). I due contesti dialogano entrambi, ma con finalità diverse, con il LDBMS. Infatti da una parte il data center ha come finalità principale la garanzia di una continuità temporale ottimale delle operazioni di archiviazione e processamento dei dati, dall'altra la "Gaia Room" è un ambiente naturale dove diverse ottiche scientifiche, per tramite del DAAS (Data Access and Analisys System) ci si attende possono avere, in tempo reale o differito, il polso dell'andamento della missione. Questo implica, dal punto di vista dell'hardware destinato al DAAS, la disponibilità di:
almeno una stazione di elaborazione destinata alle "Scientific Operations"; questa è in grado di interagire per accedere, ed ove necessario coordinare ed orientare i flussi di dati, da e verso il data center
di tanti ambienti operativi, in grado di ricevere ed analizzare dati provenienti dal LDB, quanti sono i sottosistemi da monitorare, in prima istanza: GSR, AIM e BAM.
Ma, per concretizzare quanto sopra, occorre che questi contesti siano in grado di manipolare i dati, a vario livello e con diverse finalità, presenti nell'infrastruttura. Questo pone l'interrogativo di quale sia il livello dell'infrastruttura sul quale è ammissibile possano agire, ed ove questo sia possibile, con che livello di priorità, attori e ScOM.
Questi aspetti sono indagati nelle sezioni successive, ma a questo livello si può affermare che si ritiene fondamentale la disponibilità di canali e risorse hardware locali sufficienti a consentire al personale scientifico la gestione di moli di dati in un contesto adeguato al tipo e all'intervallo temporale di analisi scientifica da condurre. Nel contempo sia attuabile la condizione, in eventuale impossibilità dell'attore di essere presso il DPCT, di poter accedere da remoto alla stazione di gestione del suo sottosistema a mezzo di una sessione di accesso remoto alla work station, ovvero mediante un'interfaccia web; quest'ultima quantomeno per le procedure di monitoraggio "real time".
Questa possibilità conduce inevitabilmente a considerare, nell'approccio di un Local PDB, la disponibilità di strumenti software adeguatamente preinstallati sulle workstation e configurati per poter essere attivati e gestiti non solo secondo modalità stand-alone ma anche, e forse soprattutto, remote.
Quest'ultimo aspetto può avere implicite conseguenze sulla tipologia di licenze di cui dotare i sistemi piuttosto che la possibilità di interagire con strumenti propri, ad esempio collocati su un proprio notebook.

Moduli costituenti il sistema

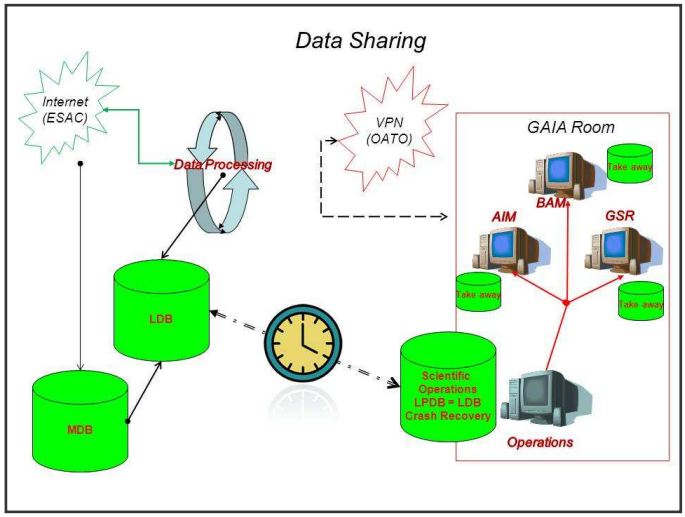
Il disegno qui riportato mostra, a grandi linee, i due contesti in cui si svolgono le attività di processamento ed analisi dei dati. Il primo risiede nel centro di calcolo di ALTEC, il secondo in quella che viene definita Gaia room. Nel centro di calcolo risiede, di fatto, il motore cui demandare tutto il carico procedurale che conduca virtualmente in modo ininterrotto dal ricevimento dei dati all'elaborazione dei medesimi, all'archiviazione finale dei dati originali e dei risultati. Il sistema è sostanzialmente automatico e la stazione preposta al monitoraggio ed alla gestione (System Monitoring Administration Subsystem) ha il solo scopo, ai fini delle Scientific Operations, di tener traccia del corretto funzionamento dell'hardware e dei processi che allocano le risorse di questo, consentendo un punto di accesso diagnostico ove si verifichino anomalie. Se gli aspetti operativi dell'ambiente del Data Center sono quindi strettamente connessi alla valutazione dello stato di salute del motore, alla Gaia Room è demandato il compito di accedere al Data Access Analysis Subsystem (DAAS), il centro preposto alla verifica che quanto il sistema sta producendo è in linea con le aspettative di contenuto scientifico. Questa semplice affermazione implica la disponibilità, in questo contesto, di strumenti diagnostici sia automatici che manuali che possano essere eseguiti, con livelli via via crescenti di efficacia e complessità, in termini di interazione con la catena di processamento ed in funzione della gravità delle eventuali anomalie che vengano riscontrate. Questa possibilità di graduare gli interventi deve tener conto, quindi, non solo dei risultati derivanti dall'analisi scientifica "ordinaria" dei dati, ma anche di elementi ancillari che forniscano altrettanto quantitativamente, in maniera preventiva, l'impatto che l'adozione di strumenti ulteriori, magari non convenzionali, può avere sul sistema. In tal senso una casistica esplorativa è da mettere a punto al più presto anche per semplice "esercizio" di interazione con il DB al fine di valutarne la percorribilità.



Inoltre se sul lato del CED deve essere possibile interagire con i processi fin anche ad arrestarli, è anche dal lato della Gaia room che deve essere possibile impartire la disposizione ultima a procedere in tal senso. In altre parole se il settore sinistro dell'immagine ha il preciso obbiettivo di rendere efficiente e continuo il processamento, sono solo i componenti della Gaia room che hanno l'obbiettivo di evitare, ovvero minimizzare, tutti quei fattori prevedibili, ma soprattutto imprevedibili, che siano dispendiosi quando non palesemente deleteri in termini di allocazione di energie computazionali e di funzionamento del DB e della Pipe Line. Inoltre la valutazione di quali modifiche, sulla base di considerazioni di ordine scientifico, apportare nel settore infrastrutturale non può che provenire dall'efficienza con cui questi aspetti sono gestiti e compresi nella "Gaia room".

Componenti e ruoli nella Gaia Room

La presenza di workstation "etichettate" nella Gaia room è legata alla esigenza di segregare, anche in questo contesto, l'ambito di interazione che i vari "attori" scientifici hanno quando operano dalla work station di pertinenza presente nella Gaia room. Questa è la postazione stand-alone appositamente prevista sia che l'attore sia fisicamente presente, sia quando da remoto acceda alle facilities attraverso, tenendo conto di quanto fin qui appurato, una VPN.



La segregazione è tuttavia solo parziale, in quanto, i vari moduli in ultima analisi accedono ad un pool di dati che si trova in una collocazione comune: il LDBMS o per traslazione parziale in un LPDMS. Con questi però interagiscono continuativamente sia i sistemi di popolamento del database stesso costituiti dal GTS, sia i processi che attingono da questo i dati necessari per la catena di riduzione e che a questo inviano i risultati dell'elaborazione. Risulterebbe quindi da escludere la possibilità che, a questo blocco operativo, possano accedere le workstation della Gaia room, in qualsiasi caso e senza prevalutazione dell'impatto, stante il rischio di introdurre in questo "ciclo virtuoso" eventi che potrebbero introdurre criticità la cui portata imprevedibile è inaccettabile.

Naturalmente diverso è, lì dove tra gli output di fasi del processamento sia stato previsto, la realizzazione di tabelle di dati da sottoporre ad analisi postuma (grafici, istogrammi, tabelle ecc.). L'affermazione implica che questi stessi dati oltre che a terminare nella collocazione naturale del LDBMS (ove previsto) vengano inoltrati ad un archivio personale che può essere anche un semplice files repository sulla propria work station o un disco take away ad hoc ad essa collegato. La struttura e la caratterizzazione di questi è auspicabilmente funzionale ad un facile reperimento dei dati (alberi di directories, codifica della tipologia del processo e del tipo di dati nel nome del file, implementazione di un DBMS) da parte dell'attore.

Questo ricollocamento è naturalmente, se condotto al limite del suo potenziale di dimensioni e spazio disco necessario, in tutto comparabile alle dimensioni del LDB, anzi, considerato lo spazio necessario per l'analisi, potrebbe essere persino maggiore, tanto che, in un approccio operativo, potrebbe coincidere con una copia di questo. Questa interpretazione del ruolo di questo spazio avrebbe il pregio di adempiere anche all'introduzione di una funzionalità di "crash recovery" pressochè istantaneo e di backup in modo implicito. Naturalmente, facendo riferimento allo schema precedente, questo connota il sistema come dotato di un Local Personal Data Base che coincide in tutto e per tutto con il Local Data Base così che la definizione del primo perde di significato. Il controllo sul corretto funzionamento di questa risorsa necessariamente pone in contiguità le due parti del Data Center e della Gaia Room in quanto la fruibilità di questa ultima risorsa è simultaneamente garantita ai processi di "data ingestion" controllati dalla catena di riduzione che la popolano e di "data retrive" primariamente esercitata dalla work station delle scientific operations e, sostanzialmente in modo paritetico, dalle CPU dei moduli AVU. Questo aspetto è però, va ribadito, subordinato alla reale disponibilità di una risorsa di "crash recovery", attualmente non prevista. In

alternativa, si può ritenere di immaginare che i dati del processo vengano indirizzati sempre ad una risorsa caratterizzata dalle caratteristiche del data model del LDMS il cui contenuto è però solo una parte ridotta dell'ammontare dei dati presenti. L'entità quantitativa e l'ambito temporale in cui i dati, che in questo caso vengono "pompati" nel repository, saranno soggetti a periodica rimozione per far posto ai dati più recenti. Nel caso della duplicazione del LDB nel PLDB, virtualmente, qualsiasi interazione con il personal database non impatta con le attività procedurali e quindi si può, al limite, immaginare una allocazione della risorsa completamente orientata alle esigenze della workstation AVU di pertinenza. L'affermazione è però solo parzialmente vera, in realtà anche in questo caso il lavoro di "new data ingestion" richiede allocazione, sia pur parziale e schedulabile, della risorsa e quindi in ultima analisi un' interazione, che ha però il pregio di essere "prevista" con il processamento dei dati. In entrambi i casi limiti sono posti anche dalle performances della rete ed in ogni caso, i limiti operativi deriveranno, strettamente, anche dalle caratteristiche dell'hardware per il quale si opti. Stabilito questo, occorre delineare cosa sia messo, e in che forma, a disposizione dell'attore scientifico. All'interno delle tabelle i dati sono strutturati secondo il data model in vigore durante quel ciclo, in linea di principio si può supporre che tale contesto non muti nel tempo e che gli standard siano quelli dei g-bin files. In realtà l'azione condotta dall'attore scientifico può determinare o un incremento del numero di dati o una loro modifica nel tempo ad opera di nuove versioni del software. In tale prospettiva, al di là dell'esigenza forte che la codifica del nome e dei formati dei contenitori deve essere concepita nell'ottica di identificare preventivamente ed univocamente i dati contenuti, si manifesta anche l'esigenza che ove questi dati non siano in qualche misura ricondotti al LDB, e quindi sull'arco di durata della missione al MDB, sia previsto dove conservarli ad uso di un accesso storico agevole per l'attore scientifico stesso. Si immagini ad esempio l'esigenza di tentare una modellizzazione ed analisi su scala biennale di un certo tipo di informazioni, nella logica di accesso al LDBMS parte di questi dati saranno probabilmente "off line" per i 3/4 e richiederanno operazioni di notevole peso subordinate alla possibilità di operare dei "restore" dai backup. Non è così nell'ottica di una scelta strategica nella quale siano le WS a consentire un locale accumulo delle informazioni, cosa praticabile , appunto, nella misura in cui si possano identificare quei parametri indispensabili a produrre anche a distanza di tempo i dati grafici a mezzo di operazioni che possono mutare nel tempo. In sostanza l'adozione di una gerarchizzazione della significatività diagnostica dei dati processati e dei risultati ottenuti che definisca cosa conservare, in che formato e dove. Analogamente possono mutare nel tempo i metodi di indagine attuabili su uguali pool di dati ed in questo caso viene avvalorata l'ipotesi di avere questi dati conservati nella forma più standard ed auto consistente possibile. Da qui anche l'opportunità di poter gestire il tutto a mezzo di s/w ampiamente condiviso nella comunità scientifica, uno dei motivi, non l'unico, per privilegiare IDL e la stesura dei dati in formato Fits.

Criteri di accesso e priorità da parte della Gaia room

L'attore che interviene con varia competenza sul ciclo di processamento deve necessariamente essere posto nelle condizioni di svolgere il suo ruolo al meglio delle possibilità strutturali del sistema, ma nel contempo deve essere inibito dal porre in atto strumenti che possano essere d'interferenza o peggio danno al sistema di processamento nel suo insieme o anche solo a parti di esso. A tal fine occorre stabilire, nei fatti, i criteri d' autorità con cui l'attore, con competenze scientifiche, interviene e il contesto su cui opera. E' oggettivamente necessario, quindi, che le richieste effettuate da questo al sistema possano essere ordinariamente esaudite, con tempi noti, se legate a aspetti procedurali previsti. Possano anche essere esaudite in modo, ancora preferibilmente automatico, ma con indicazione variabile dei tempi di risposta, quando le informazioni richieste sono inusuali, ma ancora nell'ambito del prevedibile. Divengono, finalmente, caratterizzate da un livello di priorità basso quando esulino dall'utilizzo di strumenti di analisi già previsti e richiedano quindi interazioni la cui portata ed impatto è sostanzialmente imprevedibile o prevedibilmente inaccettabile per gli effetti che avrebbe sul processamento ordinario dei dati e/o sulle risorse dell'infrastruttura.

D'altro canto è molto probabile che siano proprio le anomalie "prive" di elementi di diagnostica rapida quelle che potrebbero richiedere più pressante risposta.

L'introduzione di un criterio di "priorità" potrebbe rispondere anche a questo requisito, esso infatti implica, abitualmente, l'attribuzione di una scala di valori modulabile in funzione degli obbiettivi che si intendono raggiungere e che si traduce in diponibilità ed accessibilità ai criteri di allocazione delle risorse d'infrastruttura. Qui è proposta da 0 a 4, una graduazione che ricorda, nel principio, quella adottata per stabilire il livello di difesa da un attacco, in questo caso è l'infrastruttura che viene posta sotto stress imprevisto e con prospettive d'impatto sull'efficienza del processamento globale dei dati ineludibile. La soluzione è, quindi, quella di inserire una classificazione di priorità per l'esecuzione dei processi d'interrogazione del PLDB ovvero del LDB o dell'opportunità di chiedere il retrive di dati pregressi. L'introduzione di questo criterio implica l'adozione di un altro strumento correlato e al momento non disponibile, uno strumento che sia in condizione di "pre - valutare" tempi ed effetti della richiesta. La combinazione di questo risultato e del tipo di "forzatura" delle priorità che l'attore può esercitare determineranno gli effetti "operativi" ovvero il modo in cui, all'azione richiesta dall'attore scientifico, segue una reazione "congrua" da parte del sistema o degli attori che esercitano il controllo di infrastruttura.

In sostanza alla valutazione d'impatto segue l'attribuzione di una priorità cui viene apportato un fattore correttivo sulla base di chi richiede l'intervento. La tabella che segue delinea uno senario e non ha pretesa di avere valore di regola, quindi va considerata un esempio e nulla più.

Attore	Priorità Attribuibile				Infrastruttura
Op. Man.	0	1	2	3	Incondizionata
AVU AIM		2	3	4	Condizionata
AVU BAM		2	3	4	Condizionata
AVU GSR		2	3	4	Condizionata
IGSL			3	4	mai

Questa assunzione impedisce, di fatto, che un attore possa arrogarsi diritti di priorità tali da poter, in proprio, porre il sistema nei fatti di non operare più per i processi concorrenti di interesse di altri attori. Eccezione viene data al ruolo di op. manager nell'assunzione che a

questo venga demandata l'autorità di intervenire, in concomitanza di fattori di eccezionale ed urgente pregiudizio, attuando un più incisivo intervento fino al punto di assumere l'onere di arrestare dei processi. Un onere che deve così poter essere adempiuto a fronte di idonee informazioni, e quest'ultime devono poter essere raccolte con il necessario grado di priorità fin anche al punto di interagire amministrativamente con il LDBMS. Da quest'ultima affermazione discende che il sistema di attribuzione delle priorità è quindi, intrinsecamente, un compromesso "costruttivo" volto a bilanciare le esigenze dell'attore scientifico e quelle dell'attore infrastrutturale senza prescindere dalla possibilità di avere strumenti idonei a raccogliere tutti quei fattori che giustifichino ed avvalorino possibili interventi straordinari.
Nei paragrafi che seguono si cercherà di introdurre gli scenari "operativi" corrispondenti ai diversi tipi di interazione messa in atto al fine del reperimento dei dati così da disegnare i requisiti indispensabili al conseguimento degli obbiettivi che intende raggiungere l'attore.

Protocolli per l'implementazione di metodi di visualizzazione ordinari e nuovi

L'esigenza di utilizzare una rappresentazione visuale dei dati a valle di un'operazione di processamento avvenuta nella pipe line e, non di meno, di ottenere informazioni in corso di elaborazione dei dati nella pipeline stessa risulta dal bilanciamento di due aspetti fondamentali di uno stesso problema: quello di comprendere "se", al di là del corretto funzionamento dell'infrastruttura, i risultati parziali o finali di quanto viene prodotto sulla linea di processamento sono quanto atteso dal punto di vista scientifico e di "come" da questa informazione possa scaturire un vantaggio procedurale ovvero l'azione da intraprendere sulla pipe line, sia quest'ultima automatica, condizionata o manualmente imposta.
A tale definizione concorre: da una parte l'efficienza procedurale della pipe line che, mettendo a disposizione nel minor tempo possibile i risultati, concede maggior tempo per la loro lettura ed interpretazione e, implicitamente, la messa in atto di eventuali contromisure sul lato delle "Scientific Operations"; dall'altra un'adeguata disponibilità dello storico dei dati, un orizzonte temporale che rischia, con il procedere della missione, di rendere sempre più oneroso e complesso in termini di risorse sia h/w che s/w il reperimento dell'informazione necessaria.
Risulta per questo motivo fondamentale nella scelta delle informazioni da visualizzare l'adozione di metodi e rappresentazioni quanto più possibile completi sia in termini di diagnostica dello stato che, soprattutto, di comprensibilità delle informazioni presentate. In tal senso la soluzione del primo aspetto del problema richiede una descrizione quanto più possibile rigida delle caratteristiche che si vogliono mostrare. Ad esempio la "buona" rappresentazione grafica avrà una veste abbastanza standardizzata, con scale riportate sugli assi definite, poche informazioni chiave riportate in una legenda, alcune linee o valori di riferimento quotati a priori, una data di creazione codificata in una qualche unità definita. Tutto questo sancisce rappresentazioni statiche di dati che ben si accordano con l'aspetto di ottimizzazione dei tempi di processamento della pipe line (l'averle previste e codificate automaticamente ne quantifica tempi e modi necessari per la loro realizzazione).
Per contro, lì dove avvengano accadimenti ed anomalie, ben poco dal punto di vista diagnostico si può trarre da "rappresentazioni standardizzate" che, in maniera non prevedibile, cominciano a derogare dagli andamenti attesi. In tal senso ben si presta un approccio che consenta l'implementazione di nuove visualizzazioni di dati con redazione di grafici e file non convenzionali (ad esempio interattivi come quello IDL riportato più avanti), ma questo inevitabilmente impatterebbe ove implementato direttamente sulle attività della pipe line, con l'efficienza ed affidabilità della stessa, un prezzo inammissibile nel corso dello svolgimento della missione.
Si delineano quindi due possibili scenari:
Il primo, usuale, richiede strettamente ed obbligatoriamente interazioni standard, preventivamente certificate come compatibili con la pipe line. Per questo ogni sia pur minima variazione a quanto validato in fase di implementazione implica la verifica del prodotto in ambiente di sviluppo, la successiva validazione a mezzo test e implementazione ex novo nelle pipe line.
Di seguito, a puro titolo di esempio, per ogni sottosistema o funzionalità operante sull'infrastruttura, sono indicate le tipologie di informazioni da collezionare per dar luogo ad una visualizzazione dei dati "standard" e un modello delle specifiche cui queste rappresentazioni si ritiene debbano adempiere e che andrebbero opportunamente compilate per garantirsene la disponibilità.

Procedure per il monitoraggio scientifico standard dei sottosistemi

AVU – AIM

La procedura è finalizzata a:
Requisiti standard nel DAAS
Requisiti h/w sulla WS di Processo:
Requisiti s/w sulla WS di Processo:

AVU – BAM

La procedura è finalizzata a:
Requisiti standard nel DAAS
Requisiti h/w sulla WS di Processo:
Requisiti s/w sulla WS di Processo:

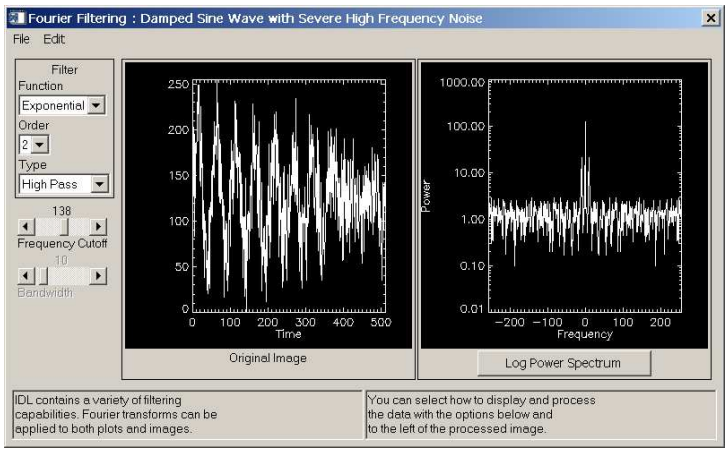
AVU - GSR

La procedura è finalizzata a:
Requisiti standard nel DAAS
Requisiti h/w sulla WS di Processo:
Requisiti s/w sulla WS di Processo:

GareX

La procedura è finalizzata a: Dare il resoconto grafico della osservabilità di Giove con Gaia fornendo una indicazione dell'evoluzione dei test di Relatività Generale attraverso la messa in evidenza degli effetti relativistici sui fotoni provenienti da stelle allorché queste transitano per moto apparente in prossimità del campo gravitazionale del pianeta.
Requisiti standard nel DAAS: Finestra statica costituita da un plot dei dati provenienti ad una certa epoca della missione desunti dall'output del processo GareX
Requisiti h/w sulla WS di Processo: Disco da 100 GB
Requisiti s/w sulla WS di Processo: JFreeChart + Gaia Lib

Il secondo, è uno scenario che è invece orientato all'esigenza di avere un'analisi estemporanea di dati con strumenti, ancorché previsti (una sorta di cassetta degli attrezzi), applicabili in maniera alternativa o cumulativa come nel caso del grafico qui di fianco proposto.



Anche per questa tipologia di scenario è necessario predisporre la definizione di requisiti che consentano l'adempimento di procedure parimenti diagnostiche e di valutazione dei dati scientifici.
La differenza per queste sarà che per le stesse non può esservi la determinazione e adempimento del prerequisito che esse siano legate ad un contesto dipendente dallo svolgimento della pipe line quindi, per tal

motivo, questi requisiti sono molto meno legati ai vincoli di efficienza e standardizzazione procedurale anzi detto. Ma implicitamente sono vincolati alla definizione a posteriori e di volta in volta puntuale di come poter reperire e interagire con i dati. Il concetto implicito in quest'approccio è che se un martello può essere ovviamente previsto come necessario per mettere un chiodo, in caso di fame e assenza di un apriscatole, va benissimo anche per aprire una scatoletta di fagioli.

Esempi di procedure per il monitoraggio scientifico non standard dei sottosistemi

AVU – AIM

Contesto
Dati ritenuti indispensabili
Formato da adottare
Requisiti h/w e s/w indispensabili sulla WS di Processo

AVU – BAM

.....

AVU - GSR

.....

GareX

La procedura è finalizzata a: *Dare il resoconto grafico dinamico della osservabilità di Giove con Gaia fornendo una indicazione dell'evoluzione del test di Relatività Generale attraverso la messa in evidenza di effetti relativistici sui fotoni provenienti da stelle allorché queste transitano per moto apparente in prossimità del campo gravitazionale del pianeta.*
Requisiti standard nel DAAS: *Finestra interattiva costituita da un plot dei dati provenienti da IGSL e dalle effemeridi del pianeta (situazione attesa) ad una certa epoca della missione e dei dati effettivamente desunti dall'output del processo GareX*
Dati ritenuti indispensabili: *Output del Processo Garex contenuti nella tabella … del LDB*
Formato da adottare: *I dati sono esportati in formato Fits nel file: GareX_data_run.fits*
Requisiti h/w sulla WS di Processo Locale: *Disco da 500 GB*
Requisiti s/w sulla WS di Processo Locale: *IDL + librerie NASA + S/W Fortran*
Requisiti attesi sulla WS Locale: *Processore i5 + 16 GB RAM – S.O. Windows 64bit*

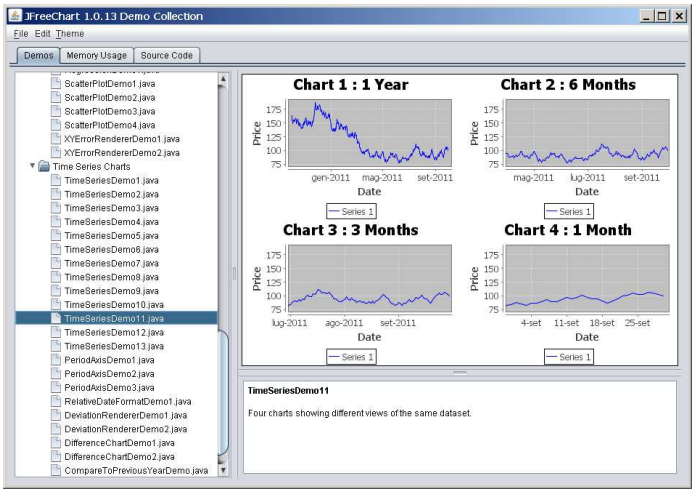
Interazioni standard

Per interazione standard con il flusso dei dati si intende una predefinita serie di informazioni ed un definito ammontare di dati che, durante l'esecuzione di uno dei sottosistemi, avvalendosi di metodi standard collezionino dati che via via vengono a rendersi disponibili o sono in qualche modo cumulati alla fine di un processo ovvero che già durante il compimento di parte di esso vengono gestiti dalla pipe line stessa e pubblicati sul DAAS in qualche forma scientificamente significativa. Queste visualizzazioni possono avere una sopravvivenza, nell'ambito del sistema, che può andare dalla durata del permanere dei dati interessati nella memoria volatile dell'infrastruttura fino ad una conservazione tombale che li colloca su disco e successivamente all'interno delle procedure di backup. In ogni caso per la loro natura diagnostica sia essa definitiva o incrementale, è fondamentale che questi tipi di dati siano sempre riconducibili al tempo in cui sono stati generati, ovvero che, ove utilizzata, la scala di tempi adottata abbia una granularità adeguata a discriminare i risultati e, finalmente, che la loro conservazione, in caso di persistenza nel LDB o nel MDB sia tale da consentirne un pronto recupero ed una agevole "rivisitazione" quantomeno a mezzo di criteri basati sul tempo.
Per la loro natura diagnostica, nel caso di ciascun sottosistema, per ogni forma di visualizzazione sia essa grafico o tabella sono indispensabili esplicite indicazioni delle unità di misura adottate, limiti di scala e tipi di scale, leggende accurate, eventuali definizioni di metodi di analisi presenti nella rappresentazione, evidenziazione della destinazione dei dati visualizzati, riferimento ad eventuali file ancillari.
La possibilità di mantenere memoria durante la rappresentazione del pool di dati che quelle visualizzazioni hanno consentito possono consentire contestualmente, all'operatore, una esplorazione "ragionata" delle curve ad esempio con il puntatore del mouse al fine di ottenere in tempo reale la lettura puntuale dei dati che quella visualizzazione hanno generato, ovvero rappresentazioni di proprietà statistiche di dettaglio. E' questo un esempio di quello che si ritiene debba intendersi come disponibilità sul DAAS di strumenti diagnostici. E' altresì auspicabile che questi stessi dati che hanno concorso alla produzione della visualizzazione possano entrare a far parte di quel pool di dati che opportunamente raccolti in un formato di esportazione, ad esempio il fits, possano essere reindirizzati all'ambiente del PDB pertinente il contesto del sotto processo (AIM, BAM, GSR primariamente, ma anche GareX o IGSL) che richieda ulteriori esami. Si ottimizza in tal modo il ricorso alle risorse di infrastruttura in previsione di quelle attività scientifiche di analisi postuma. Perché questo avvenga occorre, preventivamente, selezionare dal LDB tutti i parametri occorrenti (operazione desumibile dal contenuto dei due paragrafi precedenti) e, inoltre, in ambito di pipe line far sì che sia possibile porli a disposizione nei LPDB. Nei file trasferiti, difatti, tra tutti i dati, questi sono quelli utilizzati per la stesura delle visualizzazioni "ordinarie" e sono certamente da privilegiare in quanto consentono la riproduzione dei fenomeni evidenziati nelle interazioni avvenute durante le procedure standard. In tal senso urge la definizione del pool di strumenti di monitoraggio anche al fine di quantificare la dimensione dell'hardware presumibilmente indispensabile per le operazioni di post processo e il tipo d'ambiente operativo, ad esempio IDL (http://www.ittvis.com), Jfreechart (http://www.jfree.org) od altro, in modalità non solo stand-alone ma anche remota.
In tal senso particolare attenzione va posta, ancora, all'orizzonte temporale del supporto che la scelta di quest'ambiente garantirà nel tempo. Questo non tanto in relazione al mantenimento delle operazioni di visualizzazione standard, prevedibilmente tutte definite nelle fasi prelancio, quanto nell'opportunità di sviluppo di tools diagnostici da parte della comunità scientifica che possano essere utilmente integrati nelle fasi successive al lancio,

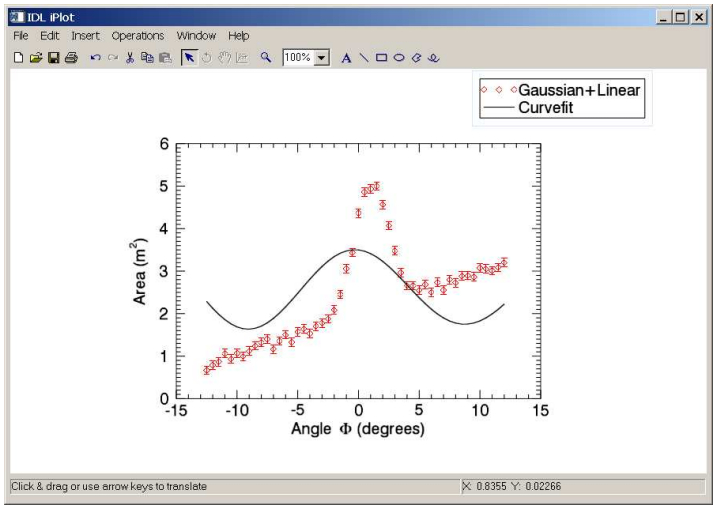
nonché nell'opportunità di potersi avvalere di procedure integrative sviluppate a supporto diagnostico per le operazioni scientifiche durante la missione.

Scientific operations

Questo contesto si ritiene abbia connotati alquanto particolari rispetto a quanto sin qui detto, infatti si colloca come linea di prima valutazione del flusso di dati visualizzati.



Al di là dei limiti oggettivi derivanti dall'impossibilità di mantenere 24 ore su 24 sotto controllo gli esiti delle varie attività di processamento, è lampante che, in realtà, di una siffatta assiduità non vi sarebbe nemmeno particolare utilità, dal momento che in ogni caso le situazioni di criticità da un punto di vista dei risultati scientifici non necessariamente possono avere riscontri che impattino sull'attività delle attività in corso sulla pipe line. Tuttavia è da quest'ambiente che possono aver inizio procedure generiche che risultino propedeutiche alla focalizzazione a vantaggio dei WP managers del team scientifico al fine di individuare quali possano essere i metodi e le contromisure da adottare. In tal senso, una opportuna valutazione della portata degli eventi può essere utile ad evitare l'attuazione di strategie, dettate dall'emotività, che possano essere infine d'intralcio all'ordinario funzionamento dei processi che avvengono sulla pipe line, così come sopravalutazioni dell'esigenza di preservare le attività di processo a qualsiasi costo possono condure ad un accumularsi di dati di fatto inutilizzabili. Per poter mediare tra le due esigenze, si ritiene utile che la parte di Operations Scientifiche possa costituire quel tramite che consente di dare un assetto proceduralmente ordinato e percorribile dal punto di vista sia infrastrutturale che scientifico senza che questo possa dar luogo a avventate allocazioni di risorse pertinenti l'infrastruttura ed analogamente scongiurando il rischio derivante da una sottovalutazione dei fenomeni che si evidenziassero nelle procedure di visualizzazione e monitoraggio.



Sembra quindi che la funzione ascrivibile a questo contesto si pone, sostanzialmente, come tramite tra l'esigenza di avere occhi umani che prendano visione con assiduità degli esiti qualitativi e quantitativi, sul piano scientifico (esempio nella prima immagine dove più rappresentazioni standard forniscono informazioni rielaborate), connessi con le attività di processamento e la possibilità che, a questo guardare, si affianchi un operare volto ad integrare i dati disponibili. Tutto questo in prima istanza, a vantaggio delle "Scientific Operations" stesse che produrrà eventualmente nuovi strumenti concepiti per una migliore e più corretta valutazione dei casi proceduralmente anomali che si vengano a creare ed, in seconda, a vantaggio delle esigenze derivanti dalla messa in opera di analisi approfondite

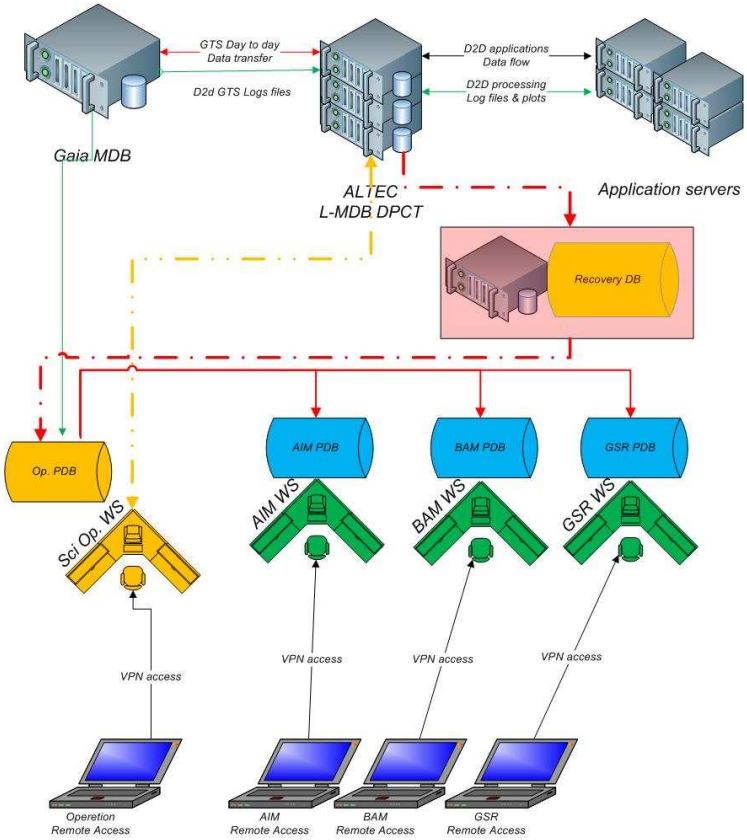
che si rendessero indispensabili alle attività del team scientifico (esempio proposto concettualmente nella seconda immagine che propone un'operazione condotta localmente ma con metodica inusuale).
Per questo motivo l'attività delle Operations scientifiche comportano la possibilità di disporre di un adeguato contingente di risorse hardware tali da garantire la conservazione per tempi adeguati di quei dati che provengono dalla linea di processamento quotidiana. Linea che origina dal corretto funzionamento della pipe line, del software su questa implementato e dalla valutazione dei risultati da questa, in modo ordinario, prodotti.
A questo si affianca, necessariamente, la disponibilità di un ambiente di analisi costituito dal software, qui riproposto essere IDL, che permetta una messa a punto di codice originale anche alle SOpM, ove necessario, idoneo allo sviluppo di nuovi metodi di indagine, anche estemporanei, adottando un linguaggio che possa essere metodologicamente di ausilio per lo scienziato esperto e competente del sottosistema nelle fasi successive e che in prospettiva possa divenire un patrimonio di codice utile, al pari di quello implementato dai WPM scientifici alle attività ausiliarie della pipe line.

Interazioni diagnostiche, predittive e dispositive.
In questo gruppo ricade la manipolazione di quei dati per i quali, per definizione, non è assolutamente proponibile una rappresentazione "a priori" e che quindi necessitano di una accessibilità e manutenzione in contesto "diverso" da quello della pipe line.
Il presupposto è che questo ambiente sia ben identificato nell'uso del software: IDL o di altro s/w indispensabile. Che il s/w implementato sulle piattaforme all'uopo indicate, sia completo in termini di pool di licenze a che si garantisca la copertura delle esigenze operative nel caso di adozione di pacchetti modulari. Che vi è una disponibilità adeguata di storage a garanzia del mantenimento e accessibilità dei dati necessari. Che il tutto sia opportunamente collegato alla rete in modo da riposizionare in quest'ambiente i dati, preferibilmente in formato Fits, provenienti dal LDBMS.

Interazioni manuali

Quando la costruzione delle informazioni necessarie a fornire una risposta ad un quesito scientifico si basa sulla disponibilità di dati presenti "solo" nel LDBMS o nel MDBMS si deve ritenere che si vuole attingere ad una tipologia di dati che per motivi temporali o di prevedibile/opportuna esigenza d'uso non sono stati spostati nella struttura del PDB sia esso cumulativo o connesso ad una particolare procedura. A quel punto l'accesso è inevitabilmente costruito attraverso un accesso estemporaneo ad esempio mediante una query sql posta tramite la consolle Oracle. Se l'accesso avviene alle tabelle del LDB o del MDB il processo entra in competizione certa con quelli che sono in corso di svolgimento e quindi l'opportunità che a questi sia dato seguito, i modi con cui questo avviene si riconducono alla disponibilità di metodi che consentano sia la pre-valutazione delle risorse che verranno allocate, sia del tempo che occorrerà attendere per il rilascio di una risposta. Ma oltre a questo aspetto "gestionale" occorre tener presente che è necessario aver la disponibilità dello strumento operativo. A questa necessità si potrebbe fornire una risposta con l'adozione della tecnica di Data Mining offerta da IDL. Il linguaggio di IDL ha dalla sua la importante proprietà di essere molto diffuso nelle comunità scientifiche Astronomiche, fattore che può condurre l'attore scientifico a poter manipolare i dati ricevuti in modo veloce e con proprietà di gestione. L'eventuale messa in evidenza, poi, che il metodo adottato va a costituire strumento di opportuna disponibilità anche per il futuro consente la messa in opera di una "cassetta degli attrezzi" il cui contenuto potrà accrescersi nel tempo.
Anche in assenza di un DBMS come metodo intermedio, tuttavia, la possibilità di avvalersi di IDL con le sue librerie astronomiche già precostituite garantisce un facile accesso a files quali i fits, i csv ecc. e quindi i metodi possono essere complementari ad una query Oracle diretta.
Naturalmente questa seconda ipotesi di strutturazione dei dati potrebbe richiedere da parte dell'utente una conoscenza almeno superficiale del linguaggio SQL e questo con il preciso scopo di muoversi con mezzi quanto più "primitivi" ma nel contempo spesso anche molto efficienti nel "brodo" di dati presente nel MDB o LDB. A questo da ultimo va aggiunto che anche per SQL come per IDL la messa a punto a fronte di particolari esigenze di procedure può dar luogo, durante lo svolgimento della missione, a una serie di "strumenti" che vanno ad arricchire il patrimonio di strumenti precedentemente indicato come la "cassetta degli attrezzi".

Conclusioni



La possibilità di concepire le attività di monitoraggio dell'avanzamento dei processi e dei risultati che da questi derivano, nonché la possibilità di interagire in modi e tempi non usuali e precostituiti con il contenuto dei DBMS o con dati da questi prodotti passa attraverso la messa a punto di alcuni strumenti che se anche piuttosto primitivi per tipologia e versatilità dovrebbero fornire alle unità operative costituenti l'Astrometric Validation Unit e in prospettiva ad un generico attore scientifico l'opportunità di accedere in modo primitivo ai dati con adeguata certezza di ottenere quanto occorrente per condurre avanti la propria indagine scientifica. L'immagine a fianco riportata sintetizza l'architettura, di massima, del sistema ed enfatizza di questo le componenti di cui occorre tener conto per una completa e funzionale realizzazione. In esso sono riportati i flussi di dati a prescindere dall'ipotesi di creazione dei Personal Data Base per i moduli di AVU. Il monitoraggio e la gestione della corretta esecuzione dei processi e dei risultati da questi derivanti, a posteriori immagazzinati nel Local Database Man. Sys sono evidenziati in verde. I flussi di dati ordinariamente prodotti vanno a popolare il PDB delle Operation per consentire l'accesso ai Work PM scientifici. In giallo ha luogo l'esecuzione di queries "on demand" che operano una produzione di dati temporanei mediante selezioni con criteri originali e popolano con le tabelle risultanti i PDB Data delle AVU units, un metodo da considerare eccezionale rispetto al flusso ordinario determinato dal "pushing" di dati previsti per lo stesso scopo dalle linee in rosso. L'operazione eseguita occasionalmente e con criteri definiti di volta in volta al fine di tutelare lo svolgimento ordinario delle attività di processo può divenire "prioritaria" rispetto ai processi ordinari in caso di esigenze propedeutiche alla risoluzione di anomalie gravi nei dati.