Risultati dal
CAESAR Hands' On Workshop for Data Providers

20/06/2023
Nodo 2000 CAESAR
Il workshop CHOW4DP (https://indico.ict.inaf.it/event/2294/) ha permesso un'interazione diretta fra i provider dei prodotti che saranno ingestiti nel prototipo CAESAR dell'archivio ASPIS e il gruppo di lavoro del Nodo 2000 del progetto CAESAR. A fronte del lavoro pregresso al workshop stesso e di quanto discusso nelle due giornate dell'evento, raccogliamo qui linee guida e indicazioni per facilitare l'inserimento dei prodotti nell'archivio. Nelle ultime pagine, riportiamo un esempio di tabella di eventi convertita da formato tabellare qualsiasi a tabella pronta per essere utilizzata.

Indicazioni per l'onboarding dei dati pronti

Tipi di files accettati per i dati:

Timeseries: JSON, CSV, FITS, netCDF
Tablesets: JSON, CSV, FITS, netCDF
Images: FITS, netCDF
Datacubes: FITS, netCDF

Contenuto minimale dei files che descrivono i dati:

1 row: column names [descrittore breve del contenuto della colonna]
1 row: Data Types [in che formato è rappresentato il dato? Integer, float, string, ...]
1 row: Null Values [come viene rappresentato il valore mancante/nullo?]
1 row: Quantity Units [in che quantità è rappresentato il dato? g, deg, m s^-2, …
Seguire le convenzioni riportate nel link sottostante:
https://www.ivoa.net/documents/Vocabularies/20230206/index.html]
1 row: Queryable columns [su quali colonne è sensato/opportuno fare una ricerca/un filtraggio dei dati?]
1 column: Time associated with data/event [prima colonna con il tempo UTC collegato alla misura/evento]
2-3 columns: Coordinates [coordinate (2D o 3D) che localizzano la misura/evento]

Data and ancillary data info preferred in SI units (CGS accepted)

Time format:

Time Scale: UTC
Time Representation: ISO 8601
times = ['1999-01-01T00:00:00.123456789', '2010-01-01T00:00:00']

Accepted Coordinates formats:

Per i dati localizzati sul Sole:

HPC Heliographic Cartesian → [HPC_Tx,HPC_Ty, HPC_distance]
HGS Heliographic Stonyhurst → [HGS_lon,HGS_lat, HGS_z]
HGC Heliographic Carrington → [HGC_lon,HGC_lat, HGC_radius]

Per i dati localizzati nell'eliosfera:

HCC Heliocentric Cartesian → [HCC_x,HCC_y, HCC_z]
HEE Heliocentric Earth Ecliptic → [HEE_lon,HEE_lat, HEE_distance]
HGS Heliographic Stonyhurst (HEEQ) → [HGS_lon,HGS_lat, HGS_distance]

Per i dati localizzati localizzati sulla Terra o near-Earth:

GEO Geographic → [GEO_lon,GEO_lat]
GEI Geocentric Earth Equatorial (Mean) → [GEI_lon,GEI_lat, GEI_distance]
GSE Geocentric Solar Ecliptic → [GSE_lon,GSE_lat, GSE_distance]
GSM Geocentric Solar Magnetic → [GSM_lon,GSM_lat, GSM_distance]
Geomagnetic + McIlwain's coordinates → [GSM_lon,GSM_lat, Lm]

Altri dati:

ICRS International Celestial Reference System → [ICRS_RA,ICRS_DEC, ICRS_distance]
Other Planetary data: as Earth Coord Systems

I data provider **devono** trasformare le loro coordinate native in uno di questi framework (aggiungendo le colonne necessarie).
Da questi framework è eventualmente possibile ottenere le coordinate che saranno standard nel database ASPIS (**HGS, GSE e ICRS**) durante il processo di Extract-Transform-Load. Ovviamente avere i dati già in coordinate HGS, GSE e ICRS è preferito.

ESEMPIO 1

	HARP_number	datetime	R_UTOV	D	X_pos	Y_pos	R_Sun
0	476	2012-09-14T03:00:03+00:00	3.3818	1	1612.9648	720.9119	1891.4001
1	476	2012-09-14T09:00:03+00:00	3.5464	1	1659.0387	731.7861	1891.2666
2	476	2012-09-14T15:00:03+00:00	3.4330	1	1700.8168	743.6094	1892.1681
3	476	2012-09-14T18:00:03+00:00	3.4068	1	1718.9973	757.7131	1892.5017
4	476	2012-09-14T21:00:03+00:00	3.5157	1	1736.0341	764.7866	1892.5091
...	...	...	...	...	...	...	...
35994	4996	2016-04-12T12:00:01+00:00	3.1618	0	710.5037	752.4746	1897.6101
35995	4996	2016-04-12T15:00:01+00:00	3.0275	0	761.5635	752.4771	1897.8841
35996	4996	2016-04-12T18:00:01+00:00	2.6275	0	812.5101	752.3945	1897.9984
35997	4996	2016-04-12T21:00:01+00:00	0.6896	0	861.5554	751.4263	1897.8456
35998	4996	2016-04-13T00:00:01+00:00	2.6462	0	909.4899	749.3489	1897.4675

35999 rows × 7 columns

Dati in formato originale. La colonna del tempo non è la prima e non è standardizzata, le coordinate spaziali sono date in pixel sul sensore e sono nelle ultime colonne. Questa tabella non è in formato accettabile per il DB.

	Time	HGS_Lon	HGS_Lat	HGS_distance	HARP_number	datetime	R_UTOV	
0	datetime	float	float	float	int	datetime	float	ir
1	-	NaN	NaN	NaN	-9999	-	NaN	-
2	UTC	deg	deg	AU	index	UTC		
3	1	1.0	1.0	1.0	1	0	1.0	
5	2012-09-14T09:00:03	93.031423	23.619544	0.004485	476	2012-09-14T09:00:03+00:00	3.5464	
...	...	...	...	...	...	...	...	
35998	2016-04-12T12:00:01	83.804831	46.324746	0.002543	4996	2016-04-12T12:00:01+00:00	3.1618	
35999	2016-04-12T15:00:01	84.219574	44.359103	0.00263	4996	2016-04-12T15:00:01+00:00	3.0275	
36000	2016-04-12T18:00:01	84.582144	42.521794	0.002721	4996	2016-04-12T18:00:01+00:00	2.6275	
36001	2016-04-12T21:00:01	84.896919	40.8321	0.002809	4996	2016-04-12T21:00:01+00:00	0.6896	
36002	2016-04-13T00:00:01	85.179033	39.23845	0.002896	4996	2016-04-13T00:00:01+00:00	2.6462	

36002 rows × 11 columns

Dati in tabella adattata alle necessità del onboarding: 4 colonne in più (coordinate spazio-temporali), 4 righe in più (descrizioni dei contenuti delle colonne). Questa tabella viene salvata in formato CSV ed è pronta per essere mandata al processo di Extract-Transform-Load per essere inserita nel DB.

ESEMPIO 2

	Time	HPC_Tx	HPC_Ty	HPC_distance	CME_num	LASCO_Start	Start_Date	Arrival_Date	PE_duration	Arrival_v	...	rel_wid	Mass
0	datetime	float	float	float	int	datetime	datetime	datetime	int	int	...	float	float
1	-	NaN	NaN	NaN	-9999	-	-	-	-9999	-9999	...	NaN	-9999
2	UTC	arcsec	arcsec	AU	index	UTC	UTC	UTC	hours	km/s	...	rad	g
3	1	1	1	1	0	1	1	1	1	1	...	1	1
5	1997-01-06T15:10:42	25.21014298570143	-3.2217961203879613	0.983319	1	1997-01-06T16:23:38.000	1997-01-07T08:35:05.600	1997-01-10T04:00:00	22.0	450	...	0.5585053606381855	580000000000000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
213	2016-11-05T04:24:05	506.60281660694386	810.7339799255432	0.991437	211	2016-11-05T04:24:05.000	2016-11-05T12:45:10.040	2016-11-10T00:00:00	16.0	360	...	0.2879793265790644	590000000000000.0
214	2017-05-23T05:00:06	0.9282420582086273	-2.2681010905074688	1.01252	212	2017-05-23T05:00:06.000	2017-05-23T10:07:51.120	2017-05-27T22:00:00	40.0	360	...	0.7853981633974483	-9999.0
215	2017-09-04T20:36:05	6.363323332940102	-11.236878221638905	1.00827	214	2017-09-04T20:36:05.000	2017-09-04T22:19:25.280	2017-09-07T20:00:00	8.0	490	...	1.1519173063162575	-9999.0
216	2017-09-06T12:24:05	31.801840180219017	-3.314008565267614	1.00786	215	2017-09-06T12:24:05.000	2017-09-06T14:00:38.840	2017-09-08T11:00:00	58.0	590	...	1.1519173063162575	-9999.0
217	2018-03-06T01:25:41	-60.00186993766876	-11.346556781440619	0.991992	216	2018-03-06T01:25:41.000	2018-03-07T02:01:12.200	2018-03-09T22:00:00	26.0	410	...	0.5585053606381855	-9999.0

217 rows × 33 columns

Definizione nome output e Scrittura del file in formato csv

SCO_Start	Start_Date	Arrival_Date	PE_duration	Arrival_v	...	rel_wid	Mass	SW_type	Bz	DST	v_r_stat	Accel.	Analyitic_w	Analyitic_gamma	filename
datetime	datetime	datetime	int	int	...	float	float	char	int	int	float	float	float	float	char
-	-	-	-9999	-9999	...	NaN	-9999	-	-9999	-9999	NaN	NaN	NaN	NaN	-
UTC	UTC	UTC	hours	km/s	...	rad	g	char	nT	nT	km/s	m/s^2	km/s	km^-1	char
1	1	1	1	1	...	1	1	1	1	1	1	1	1	1	0
:23:38.000	1997-01-07T08:35:05.600	1997-01-10T04:00:00	22.0	450	...	0.5585053606381855	580000000000000.0	S	14	-78.0	181.172	0	0	0	CME_01_param_inv.txt
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
:24:05.000	2016-11-05T12:45:10.040	2016-11-10T00:00:00	16.0	360	...	0.2879793265790644	590000000000000.0	S	11	-59.0	400.95	10.303	0	0	CME_0211_param_inv.txt
:00:06.000	2017-05-23T10:07:51.120	2017-05-27T22:00:00	40.0	360	...	0.7853981633974483	-9999.0	F	16	-122.0	307.493	0	0	0	CME_0212_param_inv.txt
:36:05.000	2017-09-04T22:19:25.280	2017-09-07T20:00:00	8.0	490	...	1.1519173063162575	-9999.0	F	10	-17.0	1497.7859999999998	0	477.57	3.1769e-07	CME_0214_param_inv.txt
:24:05.000	2017-09-06T14:00:38.840	2017-09-08T11:00:00	58.0	590	...	1.1519173063162575	-9999.0	S	7	-124.0	1654.917	-2.3911	420.03	3.1995e-08	CME_0215_param_inv.txt
:25:41.000	2018-03-07T02:01:12.200	2018-03-09T22:00:00	26.0	410	...	0.5585053606381855	-9999.0	S	14	-39.0	171.929	0	0	0	CME_0216_param_inv.txt

ESEMPIO 3

	Time	Quality	X_cen	Y_cen	R_sun	min	Max	mean	std	B0	Filename
0	datetime	float	int	int	int	float	float	float	float	float	string
1	-	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-
2	UTC	0-1	pixel	pixel	pixel	Gauss	Gauss	Gauss	Gauss	deg	char
3	1	1	0	0	0	0	0	0	0	0	0
4	2022-01-01T13:00:00	1	1160	1160	580	0	3	1.15644	0.63918	-3.06264	TSST_Mag_2022-01-01T13:00:00.fits
5	2022-01-01T14:00:00	1	1160	1160	580	0	3	1.15613	0.638917	-3.0675	TSST_Mag_2022-01-01T14:00:00.fits
6	2022-01-01T15:00:00	1	1160	1160	580	0	3	1.15582	0.638636	-3.07237	TSST_Mag_2022-01-01T15:00:00.fits
7	2022-01-01T16:00:00	1	1160	1160	580	0	3	1.15551	0.638375	-3.07723	TSST_Mag_2022-01-01T16:00:00.fits
8	2022-01-01T17:00:00	1	1160	1160	580	0	3	1.1552	0.638095	-3.08208	TSST_Mag_2022-01-01T17:00:00.fits
9	2022-01-01T18:00:00	1	1160	1160	580	0	3	1.15488	0.637818	-3.08694	TSST_Mag_2022-01-01T18:00:00.fits
10	2022-01-01T19:00:00	1	1160	1160	580	0	3	1.15458	0.637545	-3.0918	TSST_Mag_2022-01-01T19:00:00.fits
11	2022-01-01T20:00:00	1	1160	1160	580	0	3	1.15425	0.637268	-3.09665	TSST_Mag_2022-01-01T20:00:00.fits
12	2022-01-01T21:00:00	1	1160	1160	580	0	3	1.15394	0.636989	-3.1015	TSST_Mag_2022-01-01T21:00:00.fits

FAQ:

Q1) Alla fine del documento è scritto che nel database sono preferiti tre tipi di coordinate (HGS, GSE e ICRS). Mi pareva di aver capito che solo le coordinate in ingresso potessero essere di uno qualunque dei tipi della lista mentre, nella fase di ETL sarebbero state tutte convertite a un tipo unico e specifico (mi pare si parlasse di HGS), ricordo male?

R1) Purtroppo anche nel DB non possiamo ridurci ad un solo set di coordinate. HGS va bene per ciò che è intorno al Sole (vicino e lontano), GSE per quello che è intorno alla Terra, ICRS per tutto il resto....

Q2) Per quanto riguarda invece la lista del contenuto minimale del documento: si parla di 2-3 colonne per le coordinate, esistono dei casi in cui si hanno solo due colonne? Nei miei dati ho solo due angoli, la terza coordinata è la distanza ed è implicita, dovrei aggiungere la terza colonna anche se costante a 1 oppure si tratta di uno di quei casi?

R2) Si tratta di uno di quei casi. Puoi mandare i dati con solo 2 colonne, poi nel processo di ETL si aggiungerà la terza. Oppure puoi essere 'gentile' e mettercela già tu ;)