Improving Analytic Performance with Teradata

Luis Cajachahua, CoE Data Science Americas

September 2020

Agenda

- SAS In-Database sobre Teradata
- Mejores Prácticas
- Demo
- Diferenciadores
- DS2 Coding
- ¿Por qué optimizar el procesamiento?

SAS In-Database sobre Teradata

Teradata y SAS, partners

- Teradata es un vendedor Global de Soluciones SAS
- La colaboración inició el año 2007, para ser los mejores en performance analítico
- Foco en la creación conjunta de productos y casos de éxito
- Más de 450 ventas conjuntas en más de 240 clientes
- Centro de Excelencia SAS - Teradata
- Colaboración regular en el Roadmap de Productos para asegurar la integración exitosa

Las ventajas de SAS + Teradata

Juntos, SAS y Teradata hacen posible resolver los desafíos analíticos más sofisticados con un ecosistema analítico y de datos totalmente integrado.

- Escale rápidamente sus análisis para adaptarse a su negocio
- Simplifique la preparación de modelos, con procesos automatizados In-Database que forman parte de las soluciones de integración entre SAS y Teradata
- Implemente rápidamente los modelos SAS Viya y SAS 9 en producción utilizando la potencia de Vantage
- Cree modelos más complejos que aprovechen más datos, más variables y confianza de que se ejecutarán más rápidamente
- Las mejores soluciones en gestión y seguridad de datos
- Soluciones optimizadas que son fáciles de configurar, incluida la nube (AWS, Azure, etc.) Configuraciones On-Premise e híbridas.

SAS Viya

teradata.
Vantage

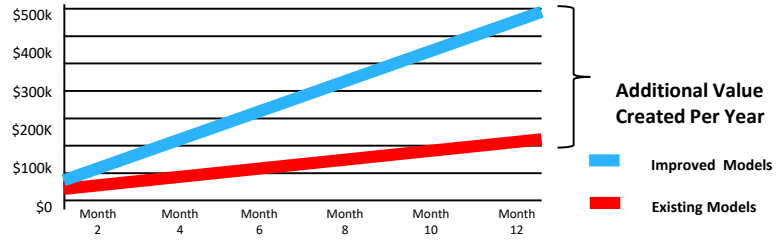
La integración de SAS & Teradata genera valor

Transformando la velocidad analítica, la disponibilidad y el rendimiento en mayores ingresos

Reducción del tiempo de desarrollo del modelo

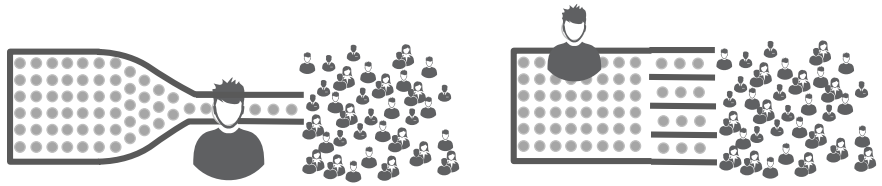
Faster analytic systems allow users to build, test and implement new models more quickly, creating additional value for the organization

Modelos con mayor performance



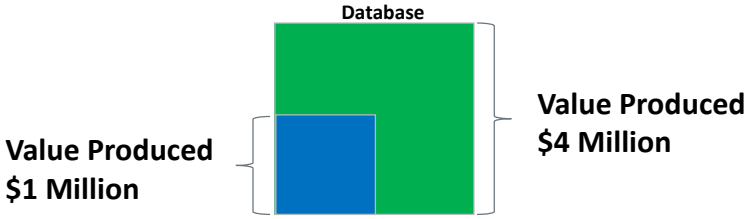
Being able to do more model testing and to update existing models to achieve optimal performance can add significant value over time

Mejora de la productividad de los analistas



Increasing the productivity of your existing data scientists reduces the need for additional headcount as well as improves model performance

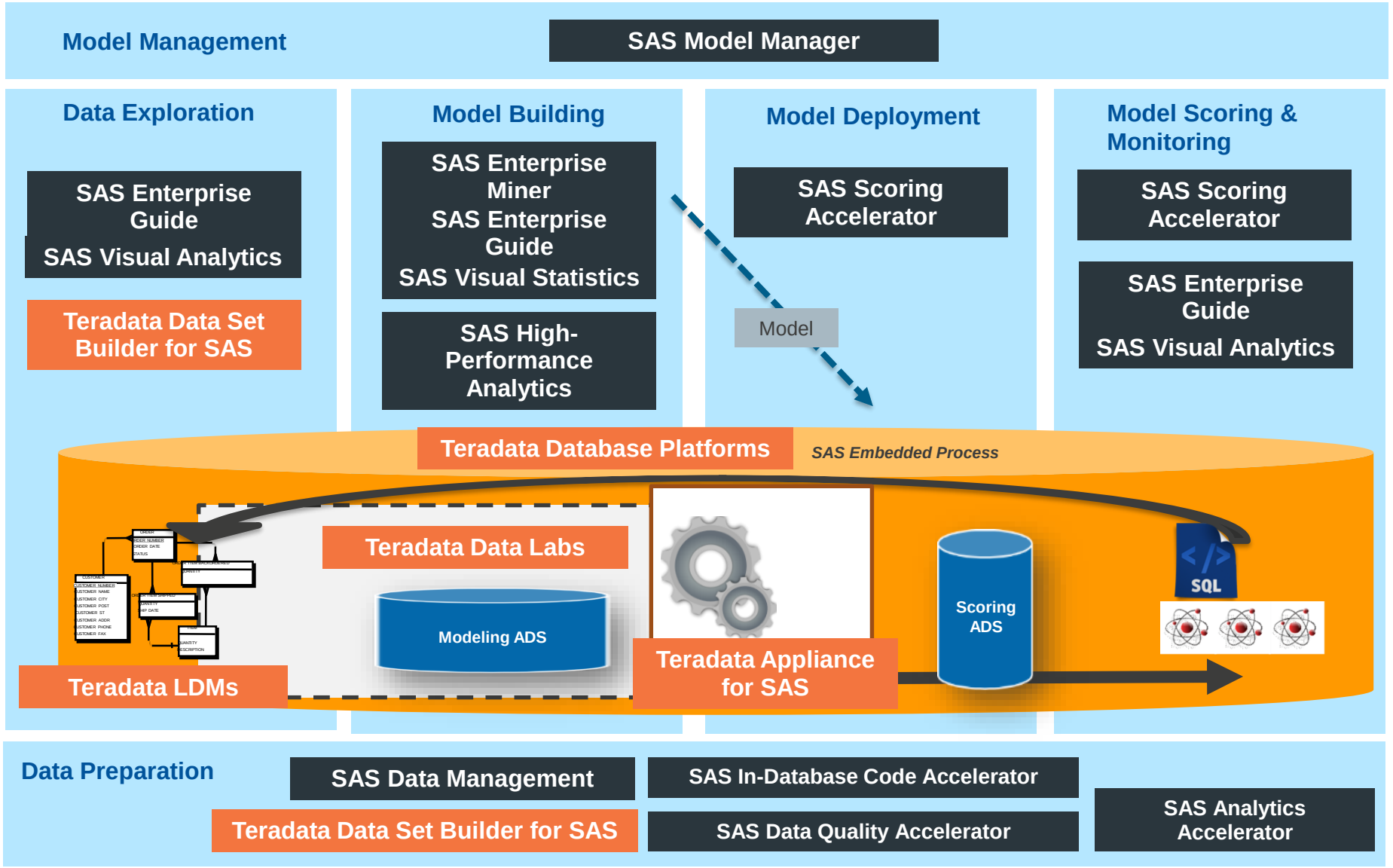
Scoring de Modelos para todos los datos



Increase the value achieved from each model by eliminating unnecessary limitations to its performance due to the sampling data

teradata.

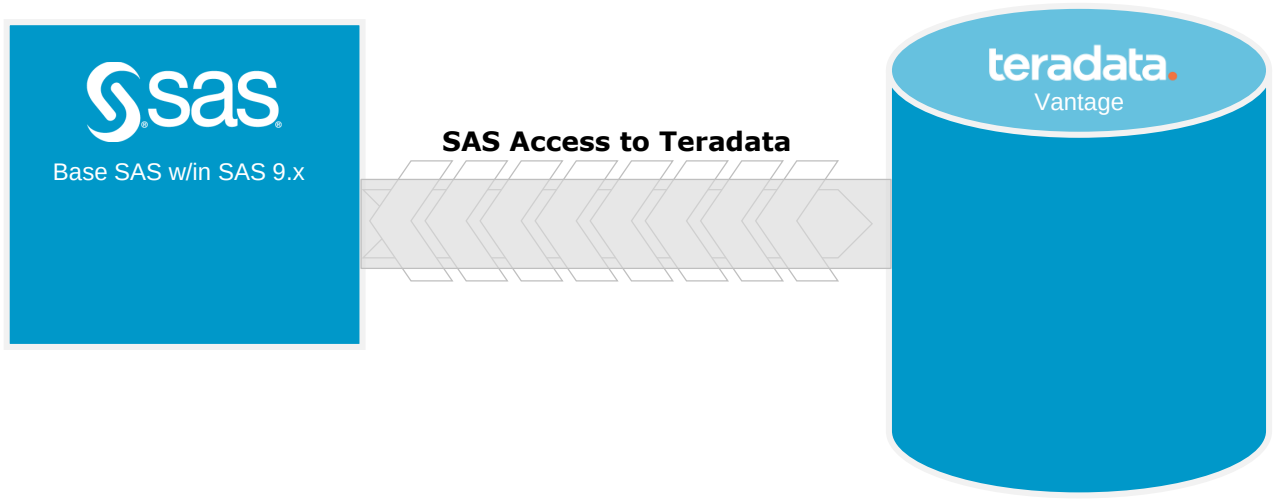
El Ecosistema de SAS 9.4



teradata.

SAS Access to Teradata

Agilizar el enlace entre los datos y sus análisis. Conectividad robusta que hace más que simplemente mover datos de un lado a otro.



Beneficios de SAS/Access to Teradata

- Permite el uso de utilidades de bases de datos por parte de los usuarios de SAS
- Velocidad de movimiento de datos.
  - Conectividad BYNET
  - Carga masiva con Teradata Fast Loader
  - Compresión de datos para almacenamiento
  - Teradata Parallel Transport
- Traduce solicitudes entre sistemas
  - Formatos de tipo de datos (mes, fecha, hoy, etc.)
  - Maneja datos especiales (nulos, valores perdidos, etc.)

teradata.

Functionalidad In-Database Nativa con SAS

SAS Access to Teradata incluye procedimientos estándar que se pueden ejecutar automáticamente in-Database para minimizar el movimiento de datos y mejorar los tiempos de modelado

IMPLICIT MODE

```
options sastrace=',,,d'
sastraceloc=saslog nostsuffix;

LIBNAME TD TERADATA DATABASE=RETAIL
SERVER='192.168.100.162' USER='dbc'
PASSWORD='dbc';
```

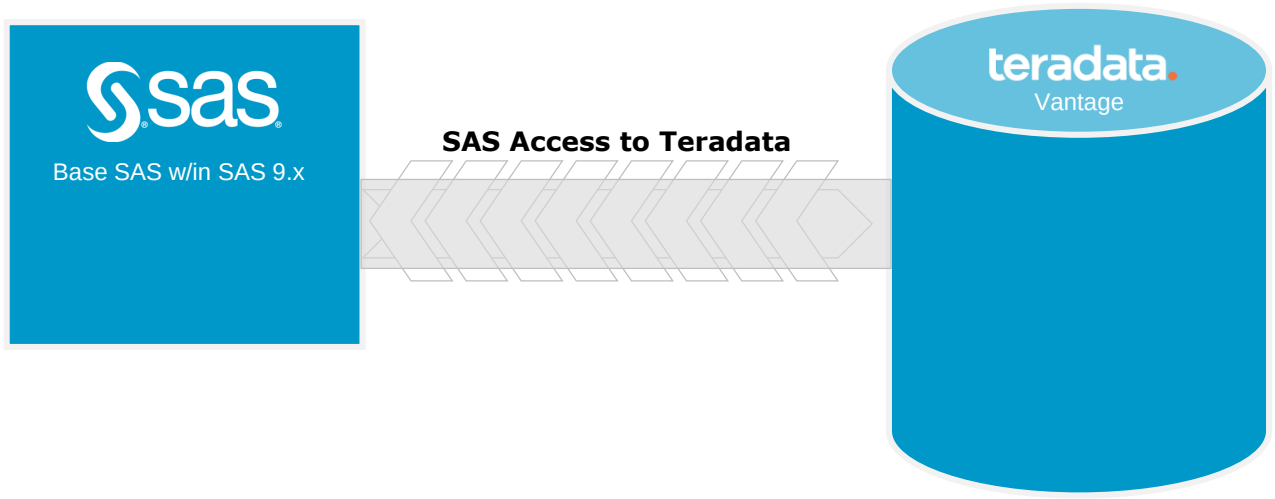
SAS/Access to Teradata Base Procedures

- PROC APPEND
- PROC COPY
- PROC DELETE
- PROC FREQ
- PROC PRINT
- PROC REPORT
- PROC SQL
- PROC TABULATE
- PROC CONTENTS
- PROC DATASETS
- PROC FORMAT
- PROC MEANS
- PROC RANK
- PROC SORT
- PROC SUMMARY

teradata.

Funciones SQL más avanzadas para SAS en Teradata

SAS PROC SQL le permite llamar a +100 funciones avanzadas de SQL que están disponibles en Vantage. Estas se inician en SAS y se ejecutan In-Database en Teradata.



EXPLICIT MODE

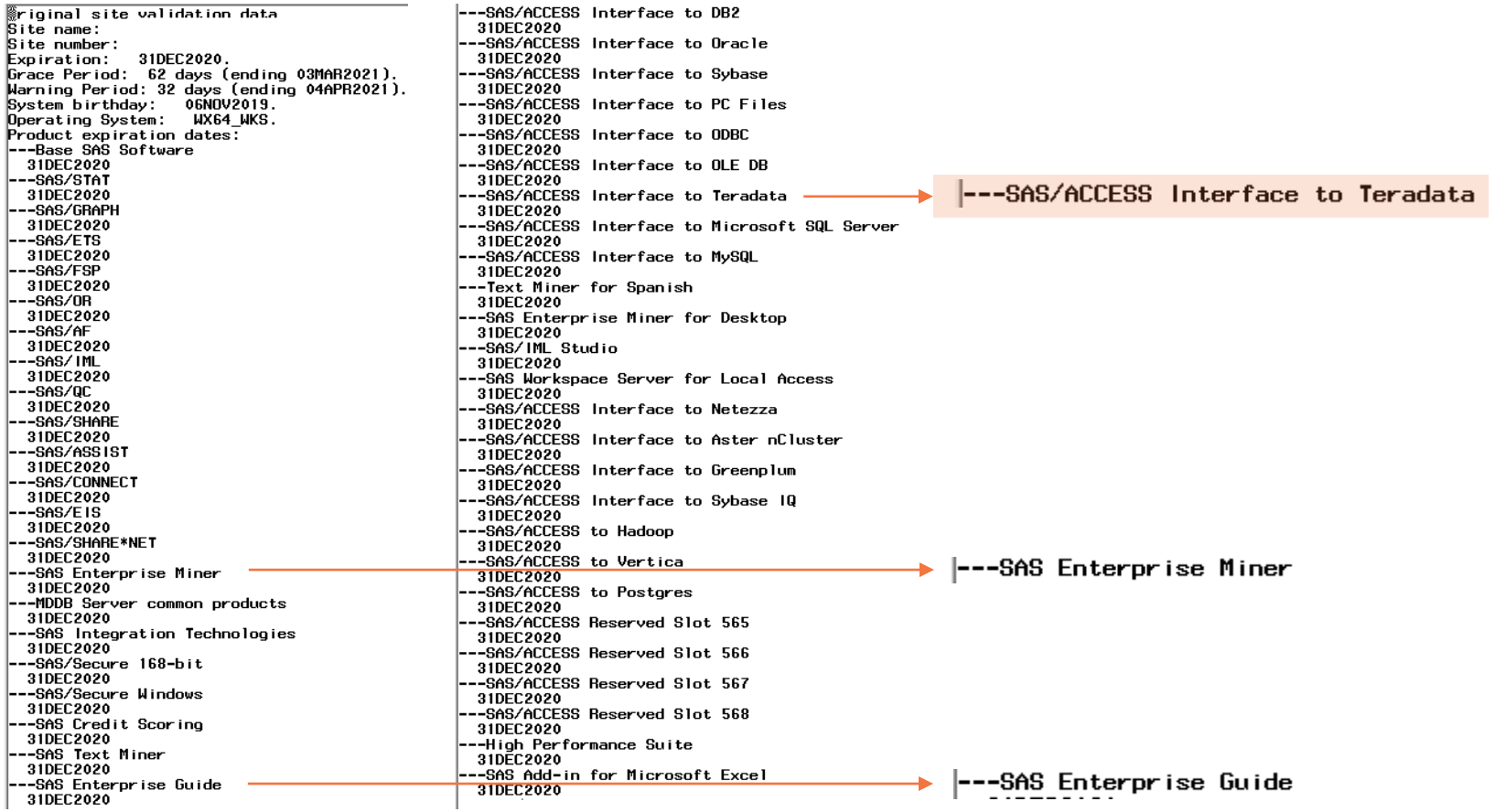
```
PROC SQL;
    Connect to Teradata ( - );
    execute (
        Teradata SQL Here
    )  by Teradata;
    disconnect from Teradata;
quit;
```

Mejores Prácticas

PROC SETINIT

¿Cómo sé si tengo SAS/Access to Teradata?

```
/* Ejecutar el script */
PROC SETINIT;
RUN;
```

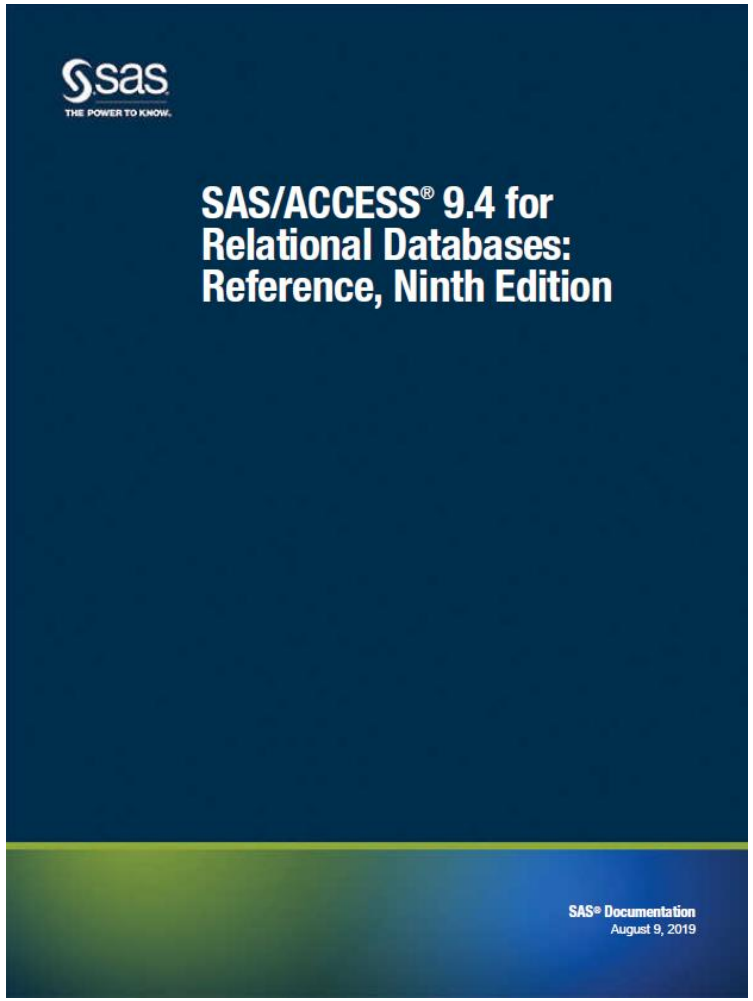


teradata.

Documentación

Acceso libre desde la web de SAS

- SAS/Access 9.4 for Relational Databases
- Westpac Presentation
- User Guide
- Accessing Teradata through SAS, common pitfalls, solutions and tips
- SAS In-Database Procedures on eBay's Teradata System Reduces Processing Time by a Factor of 4
- In-Database Procedures with Teradata: How They Work and What They Buy You



teradata.

Video Tutoriales

Existe una colección de videos sobre las funcionalidades SAS + Teradata en Youtube!

Enabling In-Database Processing with SAS/ACCESS to Teradata

SAS Software

Executing In-Database Data Quality with SAS® Data Quality Accelerator

SAS Software

Enhancing In-Database Data Preparation with SAS® Code

SAS Software

Leveraging SAS Format to Accelerate Analytics

SAS Software

Enabling Model Development In-Database with SAS® Analytics

SAS Software

Deploying and Scoring Models In-Database with SAS® Scoring

SAS Software

Exploring Millions of Rows of Data with SAS® In-Memory Analytics on

SAS Software

Managing Data and Analytics End-to-End with SAS and Teradata

SAS Software

https://www.youtube.com/watch?v=NkeyxANX_Eg&list=PL4Pq13YwpjOtsbHQTQ-jn9YdsItoeeejD

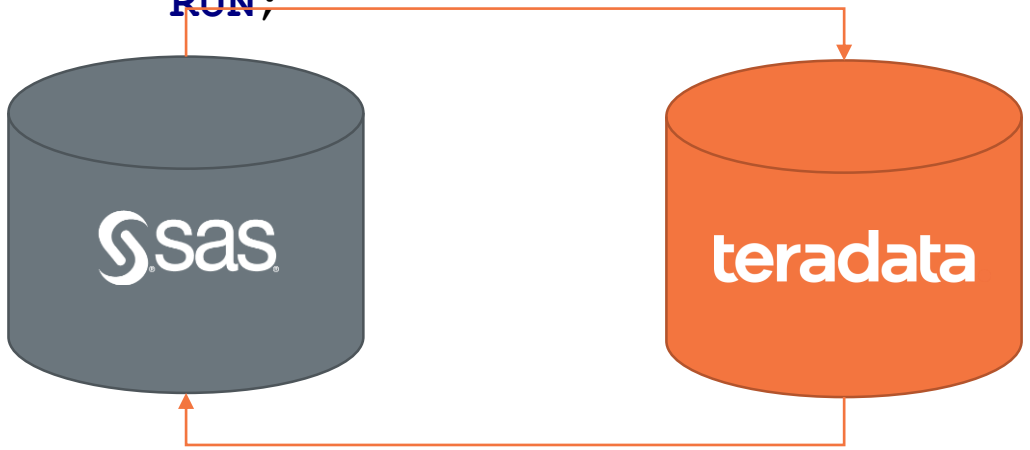
teradata.

DATA STEP

Alternativas cuando se necesite mover datos entre Teradata y SAS

De SAS A TD

```
DATA LIB.TABLA (FASTLOAD=YES TPT=YES);
SET WORK.TABLA;
RUN;
```



De TD a SAS

```
DATA WORK.TABLA;
SET LIB.TABLA (FASTEXPORT=YES);
RUN;
```

De TD A TD

```
DATA  .TABLA1;
SET  ABLA2;
RUN;
```

De TD A TD

```
CREATE TABLE BD.TABLA2 AS
(SELECT * FROM LIB.TABLA1)
WITH DATA;
```

teradata.

PROC SORT – ORDER BY

Una BD MPP no almacena datos en orden

¿Para qué?	Sugerencia
Para MERGE	Utilizar PROC SQL y JOIN
Para Informes	PROC PRINT BY
Para almacenar en SAS	DATA STEP BY

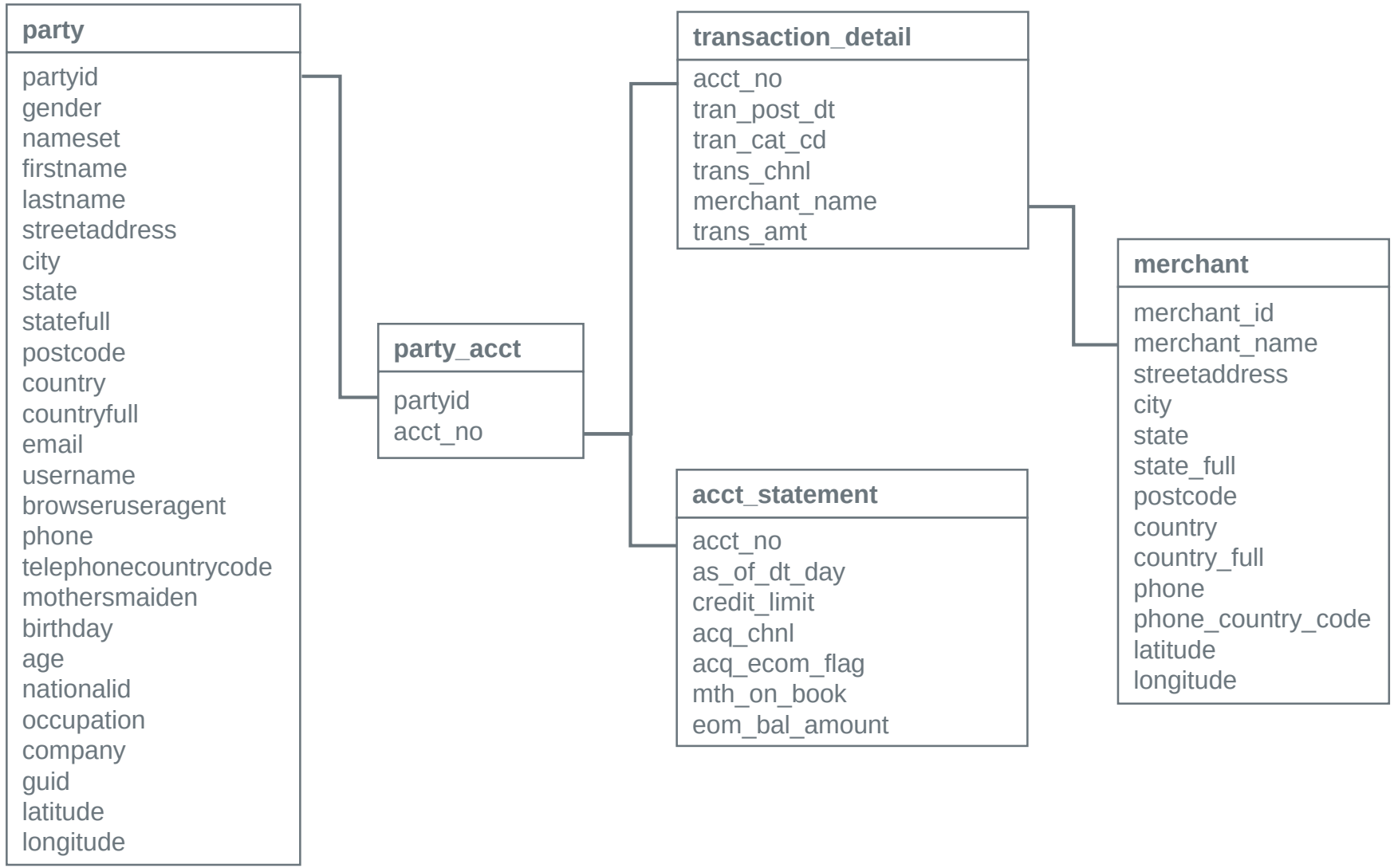
Demo

Caso: Reportes Regulatorios y Análisis de Uso de la TC

- **Entendimiento del Problema**
  - Los productos crediticios están siendo afectados por todo lo que ha sucedido en los últimos meses. El riesgo de default ha crecido para muchos clientes.
  - La estimación de provisiones debe ser más precisa. Si no, podrían dispararse.
  - Los Models de Credit Scoring indudablemente se han roto.
  - Se recomiendan análisis sencillos, alertas tempranas, etc.

- **Reportes de Ejemplo**
  - Utilización de las líneas de crédito (Sobregiro)
  - Análisis de nivel de uso y comportamiento de pago para Identificar:
    - Clientes que no utilizaban la TC y ahora empezaron a usarla mucho
    - Clientes que eran buenos pagadores pero ya no están pagando nada
    - Clientes que siguen pagando el 100%, pese a lo que viene sucediendo

Modelo de Datos Simplificado



teradata.

Reporte 1: Utilización de la Línea de Crédito

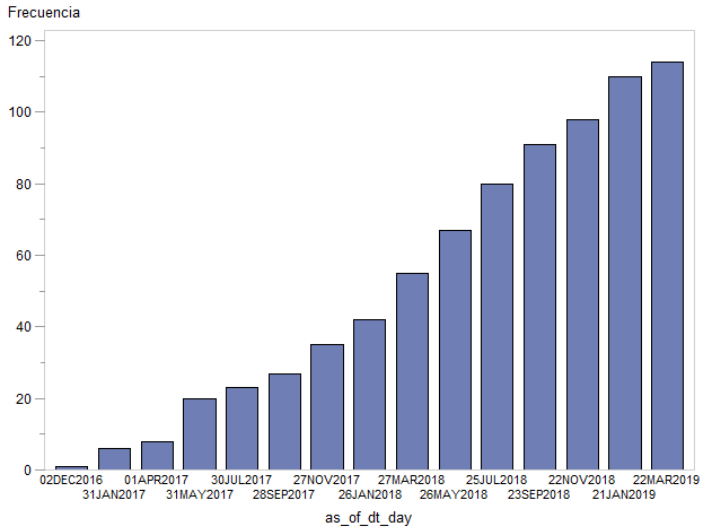
Objetivo: Identificar Clientes Sobregirados

- Se utiliza la fuente Account Statements para generar el análisis

Listado de Clientes sobregirados por cada mes

	acct_no	as_of_dt_day	credit_limit	eom_bal_amount	credit_util
1	1320	01AUG2018	1000.00	1791.00	1.791
2	1320	01SEP2018	1000.00	1791.00	1.791
3	1320	01OCT2018	1000.00	1791.00	1.791
4	1320	01NOV2018	1000.00	1791.00	1.791
5	1320	01DEC2018	1000.00	1791.00	1.791
6	1320	01JAN2019	1000.00	1791.00	1.791
7	1320	01FEB2019	1000.00	1791.00	1.791
8	1320	01MAR2019	1000.00	1791.00	1.791
9	1320	01APR2019	1000.00	1791.00	1.791
10	1647	01AUG2018	4000.00	5081.00	1.27025
11	1647	01SEP2018	4000.00	5081.00	1.27025
12	1647	01OCT2018	4000.00	5081.00	1.27025
13	1647	01NOV2018	4000.00	5081.00	1.27025
14	1647	01DEC2018	4000.00	5081.00	1.27025
15	1647	01JAN2019	4000.00	5081.00	1.27025
16	1647	01FEB2019	4000.00	5081.00	1.27025
17	1647	01MAR2019	4000.00	5081.00	1.27025
18	1647	01APR2019	4000.00	5081.00	1.27025

Cantidad de Clientes sobregirados por cada mes

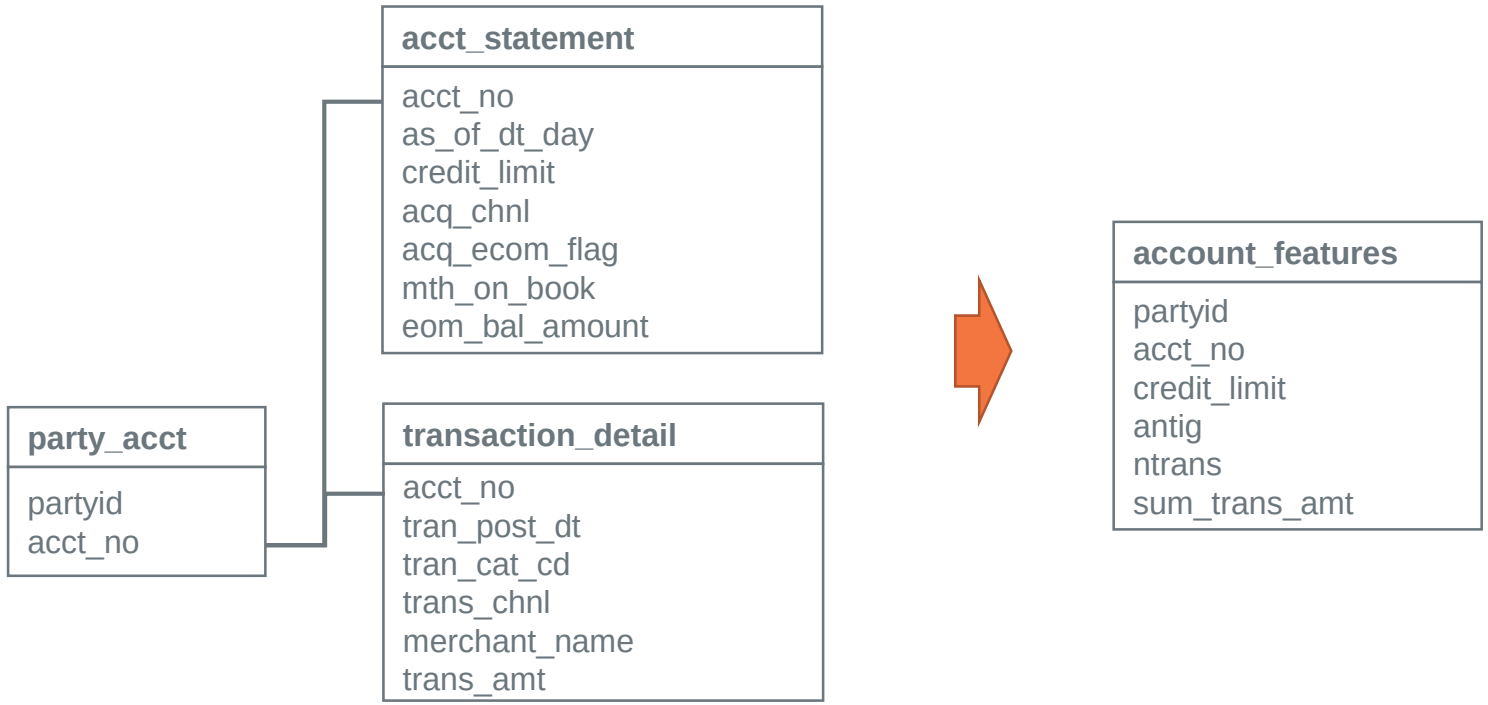


teradata.

Reporte 2: Análisis de Nivel de Uso de la TC

Objetivo: Identificar las principales métricas por cliente

- Se combinan Account Statement con Transaction Detail para calcular el nivel de uso.

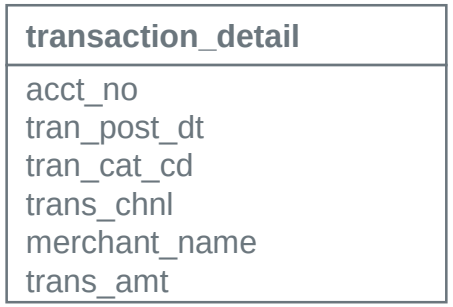


teradata.

Reporte 3: Análisis de Comportamiento de Pagos

Objetivo: Identificar nuevos segmentos de clientes

- Se utiliza la fuente Transaction Detail para generar el análisis




```
proc sql;
connect to teradata (server='192.168.100.162' user=td01 password=td01);
execute (create table td01.pagos as (
select t.acct_no, mesc, compra_amt, pago_amt, cast(pago_amt as decimal(8,2))/compra_amt ratio
from
   (select acct_no, year(tran_post_dt)*100+month(tran_post_dt) mesc, sum(trans_amt) as compra_amt
   from transaction_detail where tran_cat_cd=1 group by 1,2) as t
inner join
   (select acct_no, year(tran_post_dt)*100+month(tran_post_dt) mesp, sum(-trans_amt) as pago_amt
   from transaction_detail where tran_cat_cd=19 and trans_amt<0 group by 1,2) as p
   on t.acct_no=p.acct_no and mesc=mesp
) with data) by teradata;
execute (commit) by teradata;
quit;
```

teradata.

Diferenciadores

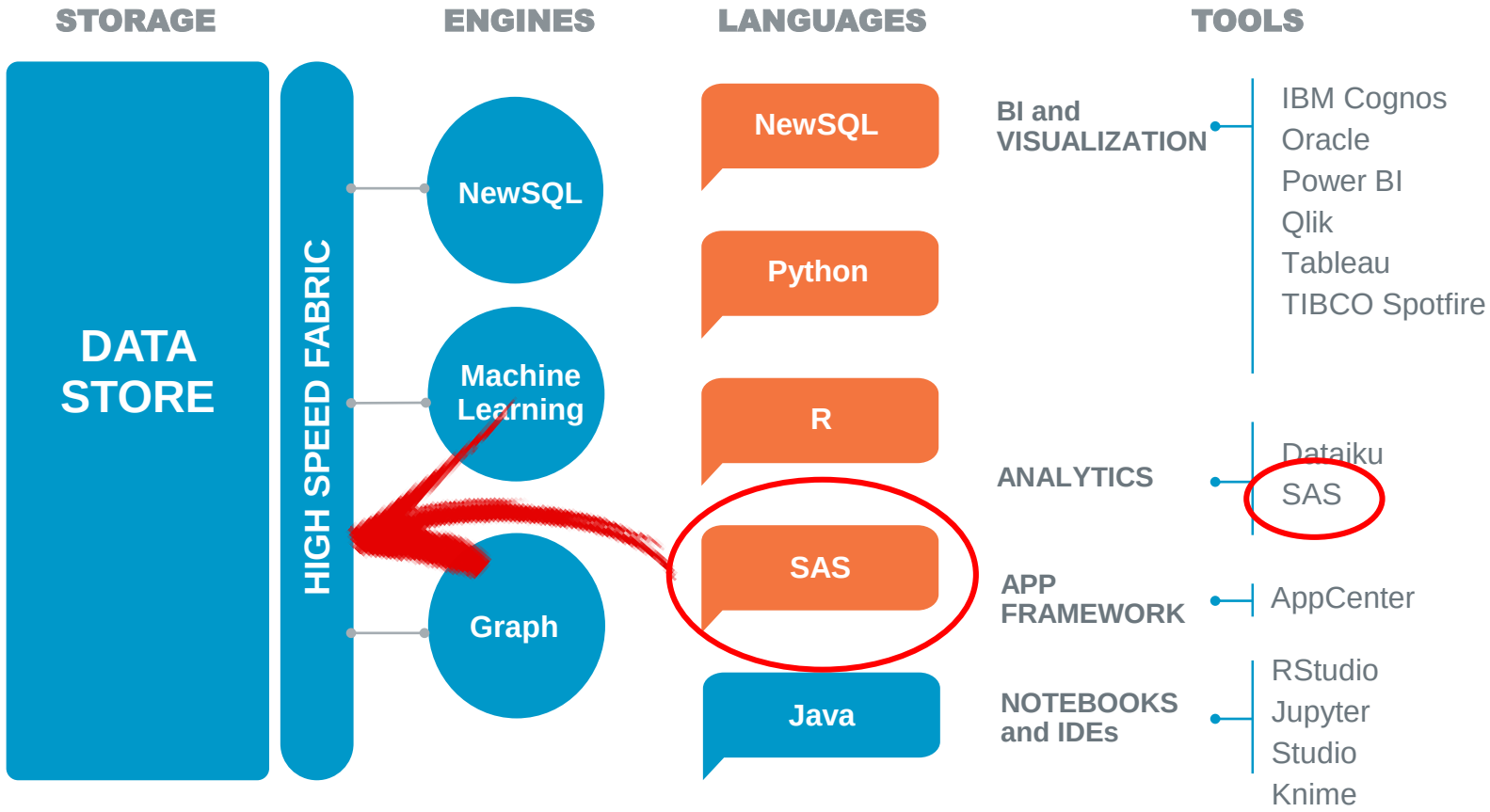
Por qué Teradata y no otras BD

Diferenciadores

- MPP (Procesamiento paralelo y arquitectura shared nothing)
- Más funcionalidades combinadas (FASTLOAD, FASTEXPORT, TPT)
- Más PROC soportados que otras BD (FREQ, MEANS, REPORT, TABULATE, etc).
- Especializada en grandes volúmenes de información y alta concurrencia.

Más que una BD, un Hub Analítico

SAS Viya & SAS 9.x pueden conectarse a Vantage para simplificar el acceso y utilizar el procesamiento directamente sobre la BD, para minimizar el movimiento de datos



teradata.

DS2 Coding

¿Qué es DS2?

Una breve introducción

- Lenguaje de Cuarta Generación, con más de 350 funciones analíticas
- Sintaxis similar a la del DATA STEP
  - Uso de DATA y SET
  - Expresiones, funciones, controles, arrays
- Influencia de SQL
  - Tipos de datos adicionales
  - SQL Embebido
- Soporta programación modular
  - Alcance y Métodos
  - Paquetes
  - Hilos

DS2 en un hilo

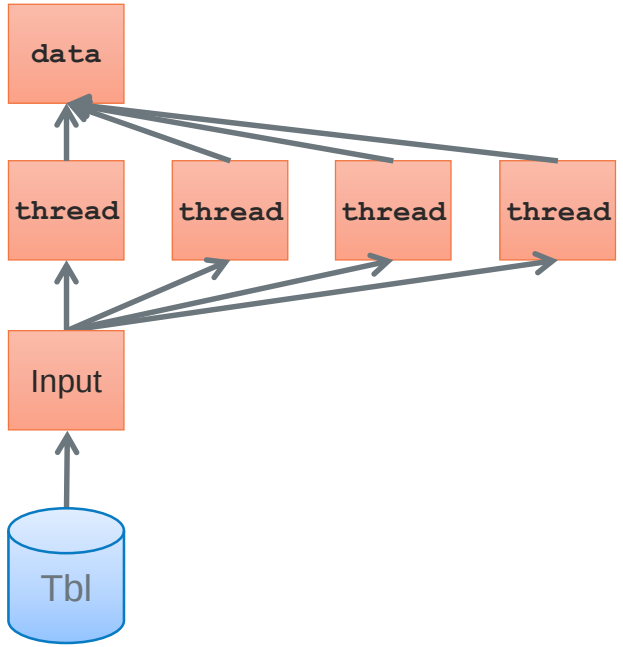
```
proc ds2;
data totals;
  method run();
    set emp_donations;
    total = sum(jan--dec);
  end;
enddata;
run; quit;
```

totals

id	jan	...	dec	total
53	100		100	1200
24	115		230	2950
87	240		45	1855
98	45		50	550

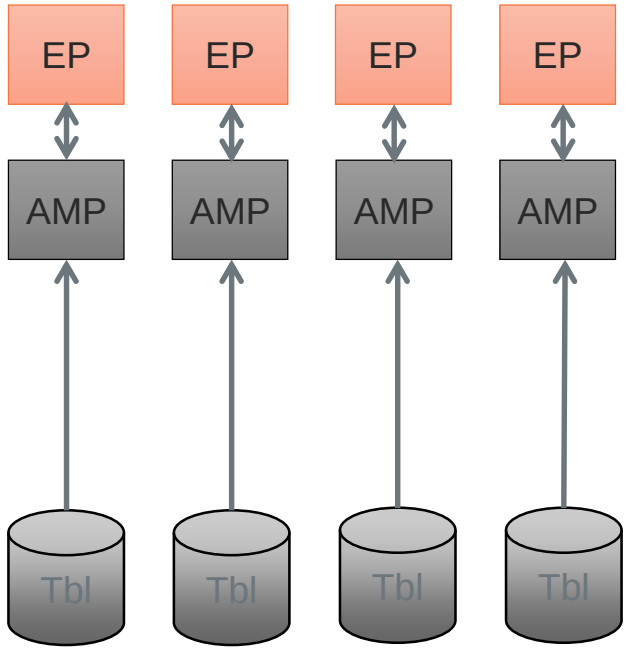
DS2 en 4 hilos – Laptop

```
proc ds2;
thread compute;
  method run();
    set emp_donations;
    total = sum(jan--dec);
  end;
endthread;

data totals;
  dcl thread compute t;
  method run();
    set from t threads=4;
  end;
enddata;
run; quit;
```



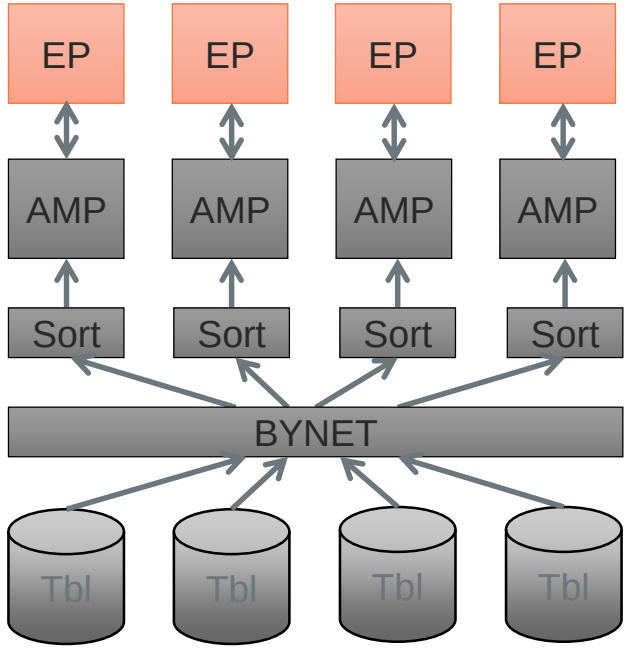
DS2 en Teradata

```
proc ds2;
thread compute;
  method run();
    set tdlib.emp_donations;
    total = sum(jan--dec);
  end;
endthread;

data tdlib.totals;
  dcl thread compute t;
  method run();
    set from t threads=4;
  end;
enddata;
run; quit;
```



Procesamiento con BY en Teradata

```
proc ds2;
thread compute;
  method run();
    set tdlib.emp_donations;
    by region;
    if first.region then total = 0;
    total + sum(jan--dec);
    if last.region then output;
  end;
endthread;

data tdlib.totals;
  dcl thread compute t;
  method run();
    set from t threads=4;
  end;
enddata;
run; quit;
```



teradata.

Code Accelerator para Teradata

- PROC DS2 lo hace automático
  - El Programa define la lógica de paralelización
  - Lee de Teradata
  - Se active con la opción DS2ACCEL=YES
  - Teradata particiona y ordena en las variables definidas en BY (opcional)
- Ambiente de Procesamiento
  - SAS Embedded Process (EP) en Teradata
  - `---SAS In-Database Code Accelerator for Teradata`
- Guía de Mejores Prácticas Disponible

¿Por qué optimizar el procesamiento?

¿Por qué optimizar el procesamiento analítico?

Mejoras reales, documentadas por los propios clientes

Tabla	Volumen	Con Otra BD	Con Teradata	Mejora
hx_surgical	129.6 G	6 Horas	50 Segundos	432X
hx_family	197.4 G	8 Horas 28 Minutos	1.45 Minutos	350X
hx_social	81.4 G	5 Horas 14 Minutos	6.46 Minutos	48X
hx_medical	240.8 G	14 Horas 27 Minutos	9.04 Minutos	95X

Tarea	Con Otra BD	Con Teradata	Mejora
Seleccionar reclamos de Medicare y Medicaid entre 2006-2013 (48 tables!)	32 Horas (1,920 Minutos)	2.1 Minutos	914X
Seleccionar reclamos de Clinical Practice Research Datalink [CPRD®] clínica, referencias, tests, inmunización, etc. de todos los años.	~6 Horas	4.5 Segundos	4800X

¿Por qué optimizar el procesamiento analítico?

Mejoras reales, documentadas por los propios clientes

Bristol-Myers Squibb

SAS Program	# of SAS Steps	SAS Only	SAS + Teradata	X Faster
1	11	1.1 Hours	1 Min	68 X
2	12	1.7 Hours	1.5 Mins	66 X
3	28	5.6 Hours	18.8 Mins	18 X
4	3,401	15.1 Hours	45.8 Mins	20 X
5	3	4.3 Days	3.8 Mins	1,648 X
6	945	9.6 Days	1.8 Hours	126 X

Teradata 2690 ~ 8 nodes ~ 192 Amps

teradata.

Thank you.

teradata.

Anexo: Pruebas de Performance y Espacio

Objetivos y Alcance

- En este caso, haremos comparaciones entre el performance de SAS contra SAS+Teradata, considerando los siguientes aspectos:
  - Espacio que ocupan los archivos finales (archivos sas7bdat versus BD).
  - Tiempos de carga de datos.
  - Tiempo de procesamiento de agrupaciones, cruces de tablas y cálculos en ambos entornos.

Los Datos

- Utilizamos una fuente de datos pública (FreddieMac), de la cual elegimos tres tablas que nos ayudarán a realizar las mediciones.
- FreddieMac es una empresa pública patrocinada por el gobierno de los Estados Unidos. Trabaja en el mercado secundario de hipotecas, ofreciendo paquetes a los inversores interesados. Por esa razón, cuentan con mucha información sobre la evolución del comportamiento de pago de los clientes finales.
- Se puede descargar la información y los diccionarios de datos de la web: https://freddiemac.embs.com/FLoan/secure/login.php?pagename=download

Prueba 1: Tiempos de Carga

Tiempo de carga de cada archivo (en segundos)

Archivo	SAS 9.4 Grid	Teradata	Mejora
historical_data1_time_Q32016.txt Tamaño: **613,716 Kb** Registros: **9,393,026** registros	31	23	1.3X
historical_data1_time_Q42016.txt Tamaño: **520,251 Kb** Registros: **7,947,665** registros	40	21	1.9X
harp_historical_data1_time.txt Tamaño: **3,817,822 Kb** Registros: **60,299,752** registros	160	96	1.7X

Prueba 2: Tamaño de los Archivos

Espacio ocupado en Kilobytes (Kb)

Archivo	SAS 9.4 Grid	Teradata	Mejora
historical_data1_time_Q32016.txt Tamaño: **613,716 Kb** Registros: **9,393,026** registros	962,048	181,820	5.3X
historical_data1_time_Q42016.txt Tamaño: **520,251 Kb** Registros: **7,947,665** registros	816,576	162,550	5X
harp_historical_data1_time.txt Tamaño: **3,817,822 Kb** Registros: **60,299,752** registros	6,126,336	1,086,430	5.7X

Prueba 3: Tiempos de Ejecución

Tiempo de ejecución de cada comando (en segundos)

Proc name	SAS Code	SAS 9.4	SAS + Teradata (SAS/Access)	Mejora
Proc FREQ	proc freq data=labdog.harp; tables period*loan_age; run;	20.28	4.78	4.2X
Proc MEANS	proc means data=labdog.harp; class period; var Act_endg_upb New_Int_rt eltv; quit;	6.92	1.67	4.1X

Proc REPORT	proc report data=labdog.harp;
	column period Act_endg_upb, SUM=Act_endg_upb_SUM;
	define period/group;
	define Act_endg_upb/analysis SUM;
	define Act_endg_upb_SUM/ format=10.2;
	title 'Report Total Amount in each month';
	rbreak after / summarize;
	run;
		5.43	1.06	5.1X

Proc TABULATE	proc tabulate data=labdog.harp; class loan_age period ; var Act_endg_upb; table loan_age, period*Act_endg_upb; title 'Amount by loan_age and period'; run;	7.15	2.00	3.6X

Prueba 3: Tiempos de Ejecución

Tiempo de ejecución de cada comando (en segundos)

Proc name	SAS Code	SAS 9.4	SAS + Teradata (SAS/Access)	Mejora
Proc RANK	proc rank data = labdog.harp ties=mean out=labdog.ranked(keep=ID_loan Period loan_age Act_endg_upb New_Int_rt eltv); by Period; var loan_age; where id_loan in ('F113Q1433500','F112Q1295837','F112Q2287280'); ranks rank_loan_age; run;	2.56	1.82	1.4X
Proc SQL	proc sql; select period, count(1) as quant, sum(Act_endg_upb) as Act_endg_upb from labdog.harp group by period; quit; run;	33.81	0.60	56.4X

Prueba 3: Tiempos de Ejecución

Tiempo de ejecución de cada comando (en segundos)

Proc name	SAS Code	SAS 9.4	SAS + Teradata (SAS/Access)	Mejora
Proc SQL (CREATE TABLE)	proc sql; create table labdog.harpres as select id_loan, period, count(1) as quant from labdog.harp where period>201512 group by id_loan, period; quit; run;	22.99	9.23	2.5X
Proc SQL (JOIN)	proc sql; select h1.period, sum(Act_endg_upb) as Act_endg_upb, count(1) as quant from labdog.harp as h1 inner join labdog.harpres as h2 on h1.id_loan=h2.id_loan and h1.period=h2.period group by h1.period; quit; run;	56.37	2.55	22.1X

teradata.