SAS FORUM ARGENTINA 2016

SAS EN HADOOP

CREANDO UN LABORATORIO BIG DATA / HADOOP / SAS

Sergio Uassouf

Líder de Práctica de
Gestión de Información e Infraestructura

BIG DATA = HADOOP + SAS

- 1024 bytes = 1 Kilobyte
- 1024 KB = 1 Megabyte
- 1024 MB = 1 Gigabyte
- 1024 GB = 1 Terabyte
- 1024 TB = 1 Petabyte
- 1024 PB = 1 Exabyte
- 1024 EB = 1 Zettabyte
- 1024 ZB = 1 Yottabyte
- 1 YB = 10^{24} bytes

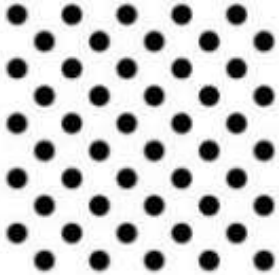
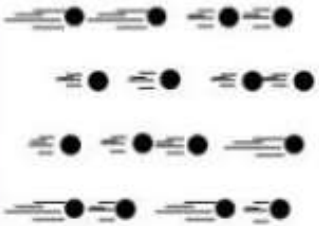
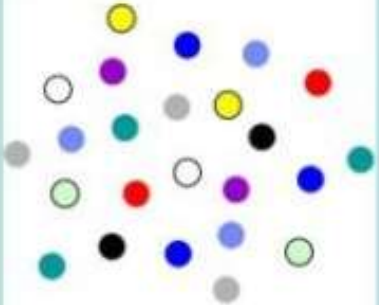
¿QUE GRANDE TIENE QUE SER LO GRANDE PARA SER GRANDE?

- Una transacción de cajero automático o home banking = 200 bytes.
- 1 TB = 5.000.000.000 de transacciones
- Red Link + Banelco ejecutan aprox. 20.000.000 de transacciones por día
- 1 TB almacena 250 días de transacciones de Link y Banelco

Entonces... ¿Qué es Big Data?

BIG DATA = HADOOP + SAS

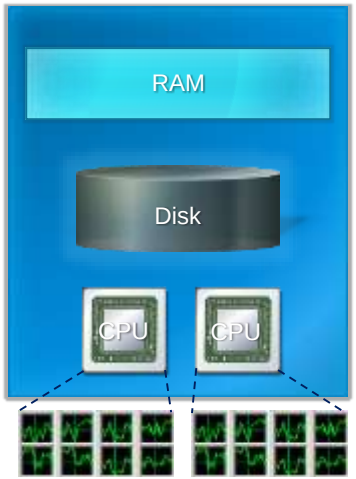
DEFINICION DE MARKETING

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

BIG DATA = HADOOP + SAS

¡¡¡ ESTO ES UN PARADIGMA !!!

- Desde los inicios de la informática un computador, ya sea personal o empresarial está compuesto de 3 componentes principales.



MEMORIA

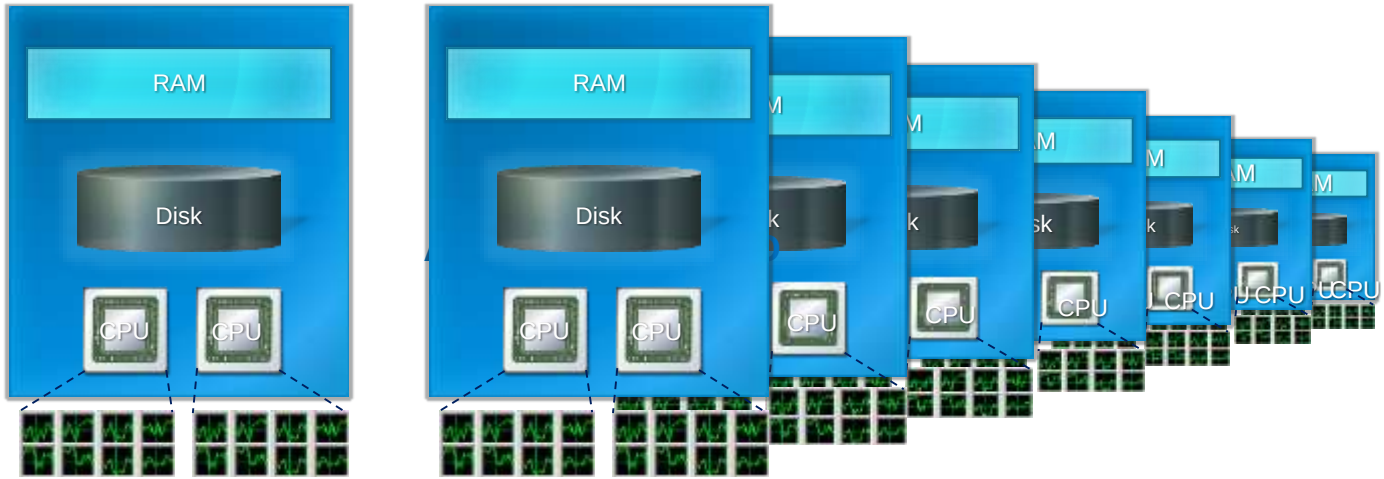
UNIDADES DE ALMACENAMIENTO

UNIDADES DE PROCESAMIENTO

BIG DATA = HADOOP + SAS

ESTO ES BIG DATA

- Desde los inicios de la informática un computador, ya sea personal o empresarial está compuesto de 3 componentes principales.



BIG DATA = PROCESAMIENTO MASIVAMENTE PARALELO

BIG DATA = HADOOP + SAS

ALMACENAR Y ANALIZAR GRANDES VOLUMENES DE INFORMACIÓN A BAJO COSTO

- Si puede almacenar mucha más información a un costo mucho menor...
- Y puede procesarla en un tiempo mucho menor.
- Entonces no necesita armar modelos tomando sólo un subconjunto de los datos...
- Y puede hacer todas las iteraciones que necesite.
- **Entonces puede almacenar y procesar la información que antes no podía**

BIG DATA = HADOOP + SAS

ALMACENAR Y ANALIZAR GRANDES VOLUMENES DE INFORMACIÓN A BAJO COSTO

TODOS LOS CALL DETAIL RECORDS

TODAS LAS TRANSACCIONES

TODAS LAS SECUENCIAS DE SITIOS WEB

TODAS LAS CONVERSACIONES DE LOS CALL CENTERS

Y ANALIZARLOS EN SU TOTALIDAD...

EJECUTANDO TODAS LAS ITERACIONES QUE NECESITE...

A MUY BAJO COSTO RELATIVO

BIG DATA = HADOOP + SAS

COMPONENTES DEL "ECOSISTEMA"

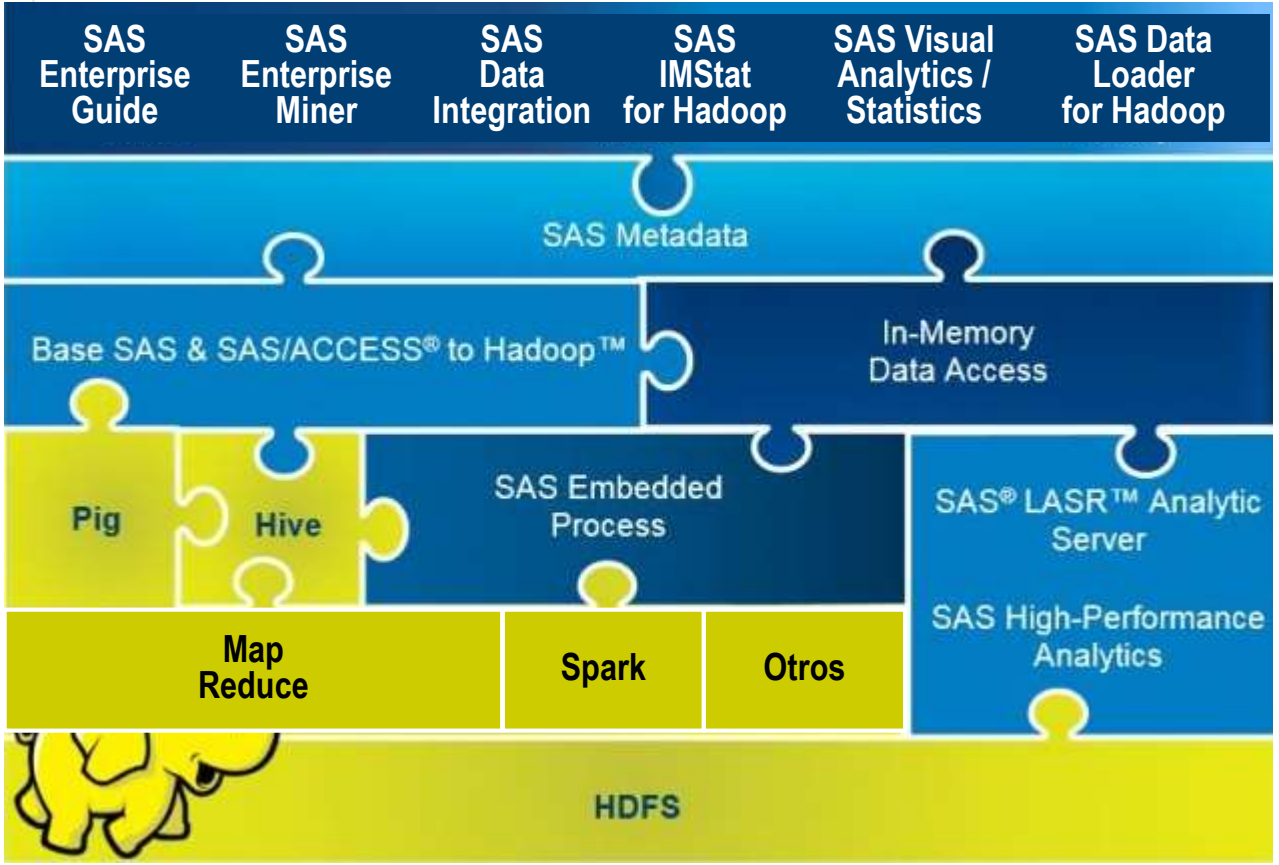
INTERFAZ DE USUARIO

METADATOS

ACCESO A DATOS

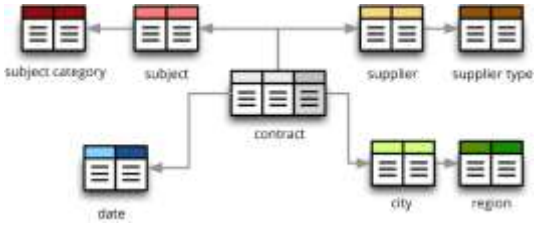
PROCESAMIENTO DE DATOS

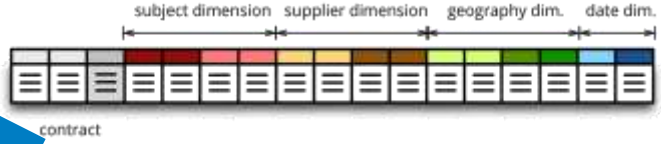
FILE SYSTEM



BIG DATA = HADOOP + SAS

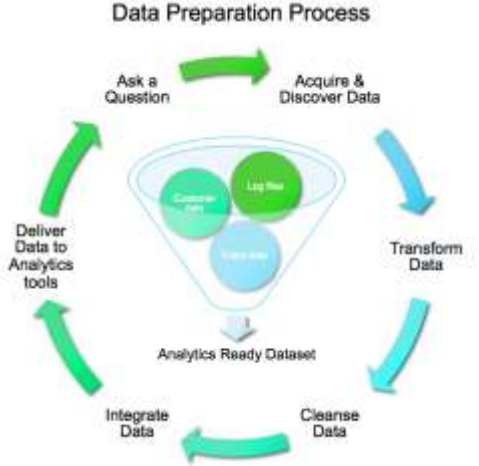
PREPARACIÓN Y ANÁLISIS DE DATOS

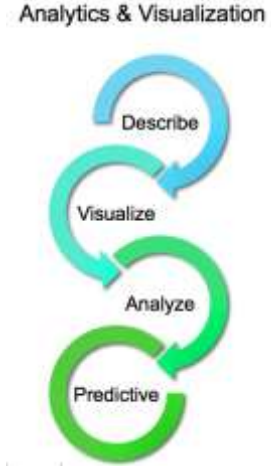




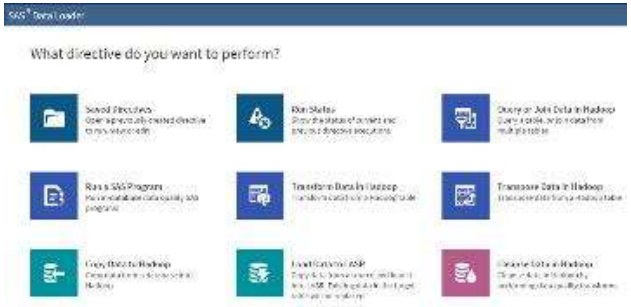
BIG DATA = HADOOP + SAS

PREPARACIÓN DE DATOS "IN-DATABASE" ESTADÍSTICA Y MINERÍA DE DATOS "IN-MEMORY"





SAS Data Loader for Hadoop



Visual Analytics / Statistics

HADOOP | PROCESAMIENTO "IN-DATABASE"

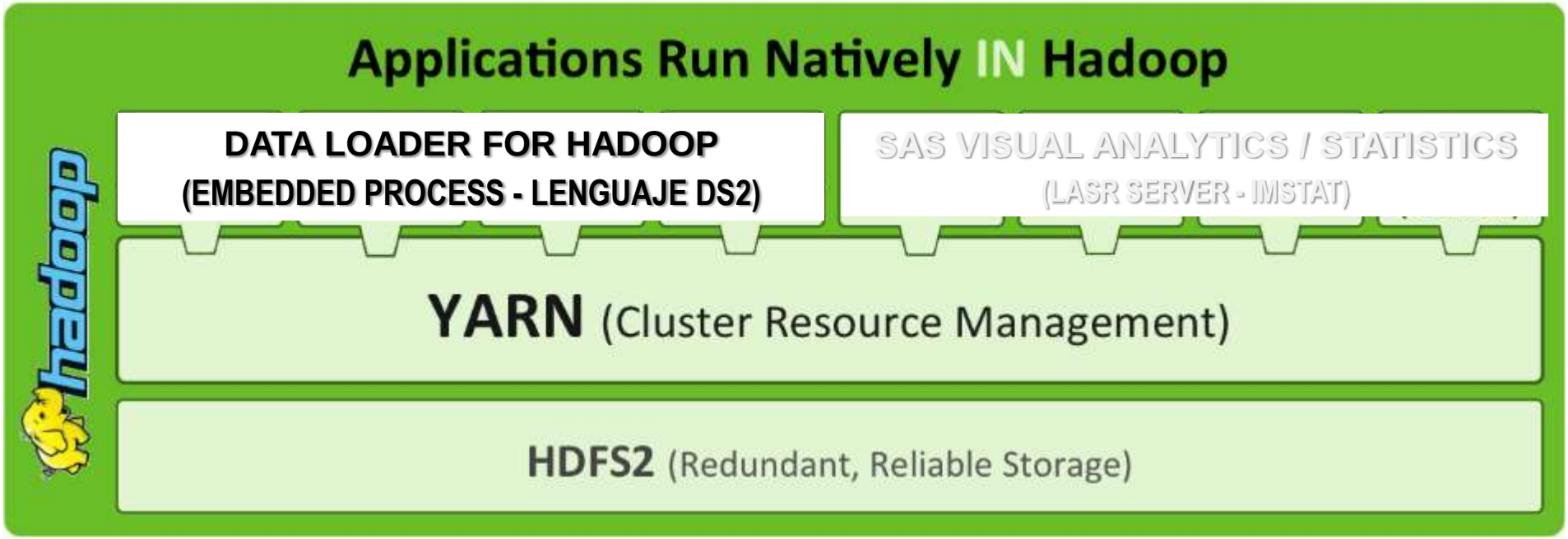


Gráfico tomado de Hortonworks

HADOOP PROCESAMIENTO "IN-DATABASE"

- Embedded Process: Traduce código de alto nivel a Map-Reduce.

```
proc ds2 ;
/* thread ~ eqiv to a mapper */
   thread map_program;
   method run(); set dbmslib.intab;
   /* program st      s */
   end; endthre
   /* program
   data hdf.dat
dcl thread map_p        ap_pgm; method
run();
set from map_pgm threads=N;
/* reduce steps */ end; enddata;
run; quit;
```

1. Integración
2. Preparación
3. Calidad
4. Scoring

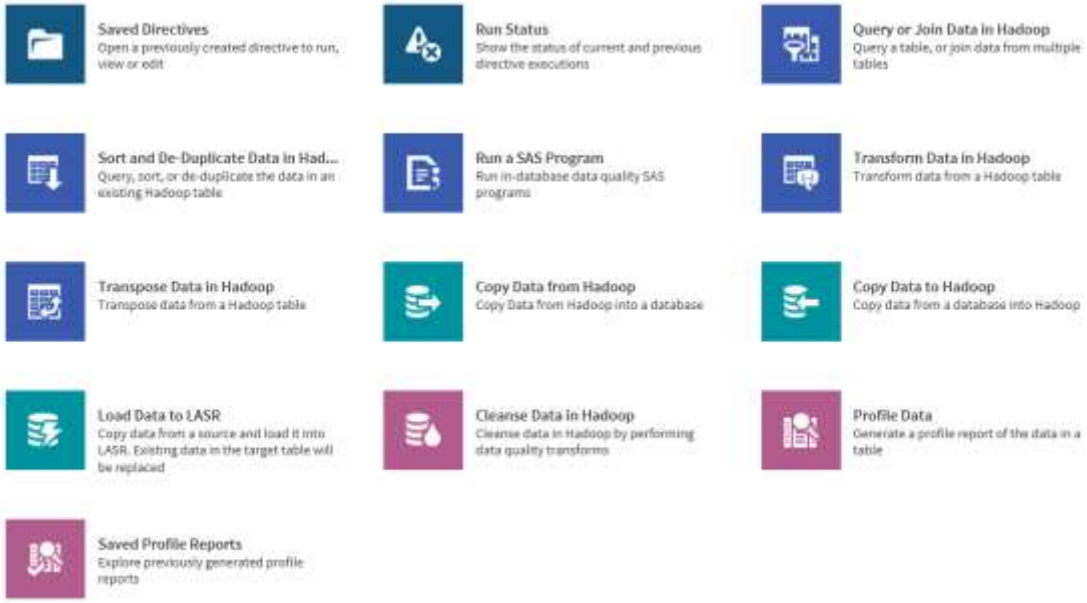
DATA LOADER FOR HADOOP

IDEAL PARA PREPARACIÓN DE TABLAS ANALITICAS

CODIGO ALTO NIVEL EN HADOOP

```
proc ds2 ;
/* thread ~ eqiv to a mapper */
   thread map_program;
   method run(); set dbmslib.intab;
   /* program statements */
   end; endthread; run;
   /* program wrapper */
   data hdf.data_reduced;
dcl thread map_program map_pgm; method
run();
set from map_pgm threads=N;
/* reduce steps */ end; enddata;
run; quit;
```

DATA LOADER FOR HADOOP



DATA LOADER FOR HADOOP

MENÚ PRINCIPAL DE DIRECTIVAS

Directivas guardadas
Abrir una directiva ya creada para ejecutar, ver o editar

Estado de ejecución
Mostrar el estado de las ejecuciones de directivas anteriores y actuales

Consultar o unir datos ...
Consultar una tabla o unir datos de múltiples tablas

Ordenar y deduplicar d...
Consultar, ordenar o deduplicar datos de una tabla Hadoop existente

Ejecutar programas SAS
Ejecutar programas SAS de calidad de datos en la base de datos

Transformar datos en H...
Transformar datos de una tabla Hadoop

Transponer datos en H...
Transponer datos de una tabla Hadoop

Copiar datos desde Ha...
Copiar datos desde Hadoop en una base de datos

Copiar datos en Hadoop
Copiar datos de una base de datos en Hadoop

Cargar datos en LASR
Copiar datos de una fuente de datos y cargarlos en LASR. Se reemplazarán los datos exis...

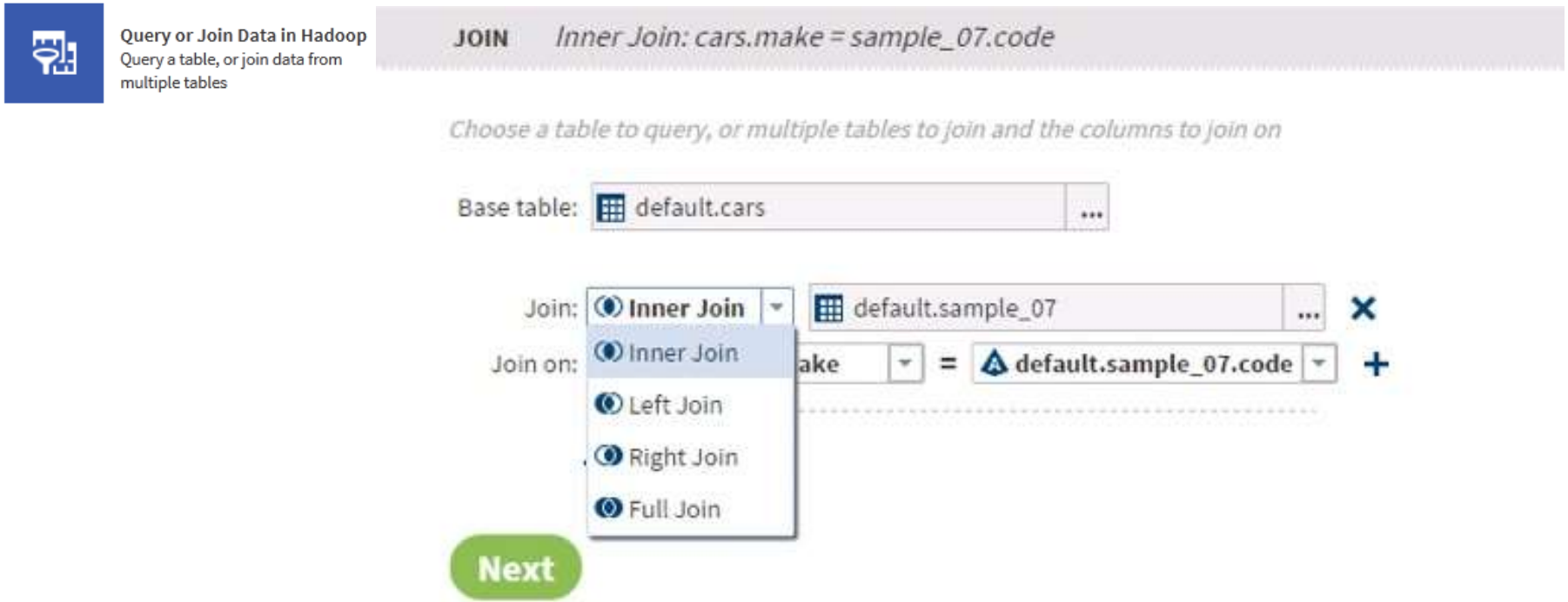
Limpiar datos en Hadoop
Limpiar datos en Hadoop realizando transformaciones de calidad de datos

Datos del perfil
Generar un informe de perfil de los datos en una tabla

Informes de perfil guar...
Explorar informes de perfil previamente generados

DATA LOADER FOR HADOOP

QUERIES Y JOINS



DATA LOADER FOR HADOOP

MENÚ PRINCIPAL DE DIRECTIVAS

Directivas guardadas
Abrir una directiva ya creada para ejecutar, ver o editar

Estado de ejecución
Mostrar el estado de las ejecuciones de directivas anteriores y actuales

Consultar o unir datos ...
Consultar una tabla o unir datos de múltiples tablas

Ordenar y deduplicar d...
Consultar, ordenar o deduplicar datos de una tabla Hadoop existente

Ejecutar programas SAS
Ejecutar programas SAS de calidad de datos en la base de datos

Transformar datos en H...
Transformar datos de una tabla Hadoop

Transponer datos en H...
Transponer datos de una tabla Hadoop

Copiar datos desde Ha...
Copiar datos desde Hadoop en una base de datos

Copiar datos en Hadoop
Copiar datos de una base de datos en Hadoop

Cargar datos en LASR
Copiar datos de una fuente de datos y cargarlos en LASR. Se reemplazarán los datos exis...

Limpiar datos en Hadoop
Limpiar datos en Hadoop realizando transformaciones de calidad de datos

Datos del perfil
Generar un informe de perfil de los datos en una tabla

Informes de perfil guar...
Explorar informes de perfil previamente generados

DATA LOADER FOR HADOOP

MENÚ DATA CLEANSING

Limpiar datos en Hadoop
Limpiar datos en Hadoop realizando transformaciones de calidad de datos

Filter Data
Select the rows of data to include

Generate Match Codes
Create match codes for selected values in the table

Identification Analysis
Identify the semantic data type of text in selected columns

Manage Columns
Select the columns to include

Parse Data
Select the column, Definition, and Token you want to apply, and enter a name for the new column

Standardize Data
Apply data standards to selected columns

Summarize Rows
Create a new row with data summarized in selected columns

HADOOP PROCESAMIENTO "IN-MEMORY"

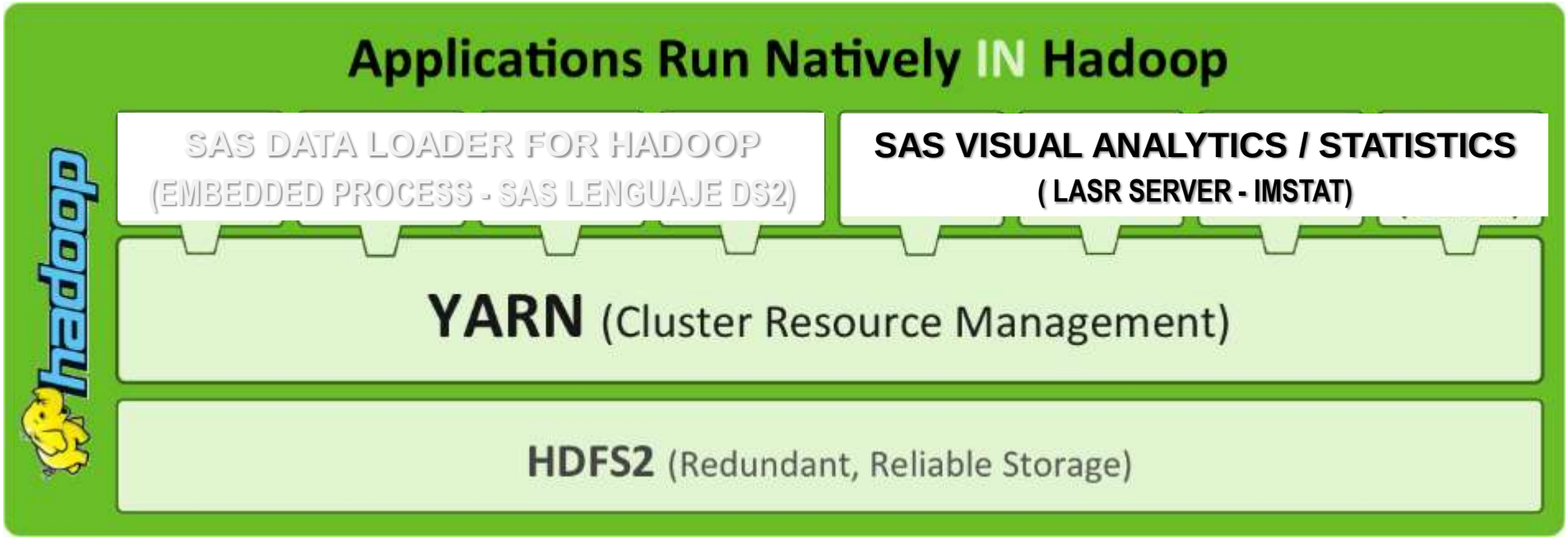


Gráfico tomado de Hortonworks

SAS LASR ANALYTICS SERVER

Data Manipulation
- SAS Data Step
- BALANCE
- COLUMINFO
- COMPUTE
- DELETEROWS
- DISTINCT
- DROPTABLE
- FETCH
- GROUPBY
- PARTITION
- PROMOTE
- PURGETEMPTABLES
- SET
- TABLE
- UPDATE

Data Exploration/ Visualization
- BOXPLOT
- CORR
- CROSSTAB
- CONTOURPLOT
- DISTRIBUTIONINFO
- FREQUENCY
- HISTOGRAM
- KDE
- REPLAY
- SUMMARY

Predictive Modeling
- DECISIONTREE
- FORECAST
- GENMODEL
- GLM
- RANDOMWOODS
- ASSESSMENT

Descriptive Modeling
- CLUSTER
- CLUSTER TF-IDF
- ASSOCIATIONS
- SVD

Recommender
- CLUSTER
- KNN
- ASSOCIATIONS
- SVD

Text Analytics
- PARSING
- SVD

Miscellaneous
- EXTERNAL (C API)
- FREE
- SAVE
- STORE

Deployment
- SCORE

SAS LASR ANALYTICS SERVER

INTERFAZ PARA PROGRAMACIÓN POR CÓDIGO E INTERFAZ PARA ANÁLISIS EN MODO GRAFICO / VISUAL

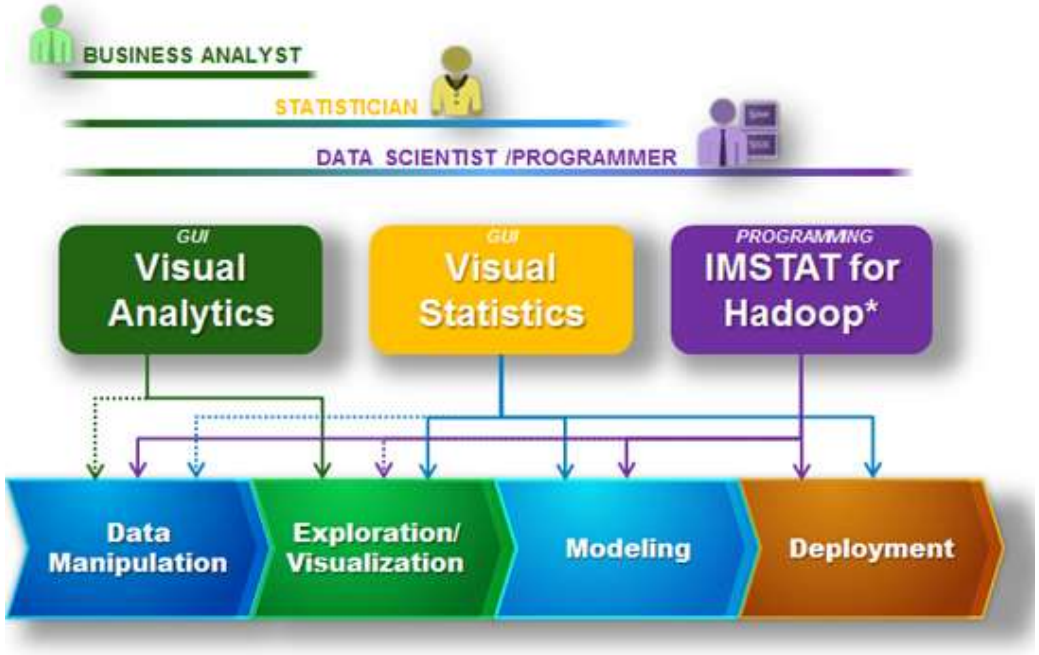
IN-MEMORY STATISTICS

```
data example.iris;
      set sashelp.iris;
run;
proc imstat data=example.iris;
      corr;
quit;
```

Pairwise Correlations for Table WORK.IRIS

Column	Row	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	1	1.0000	-0.1176	0.8718	0.8179
SepalWidth	2	-0.1176	1.0000	-0.4284	-0.3661
PetalLength	3	0.8718	-0.4284	1.0000	0.9629
PetalWidth	4	0.8179	-0.3661	0.9629	1.0000

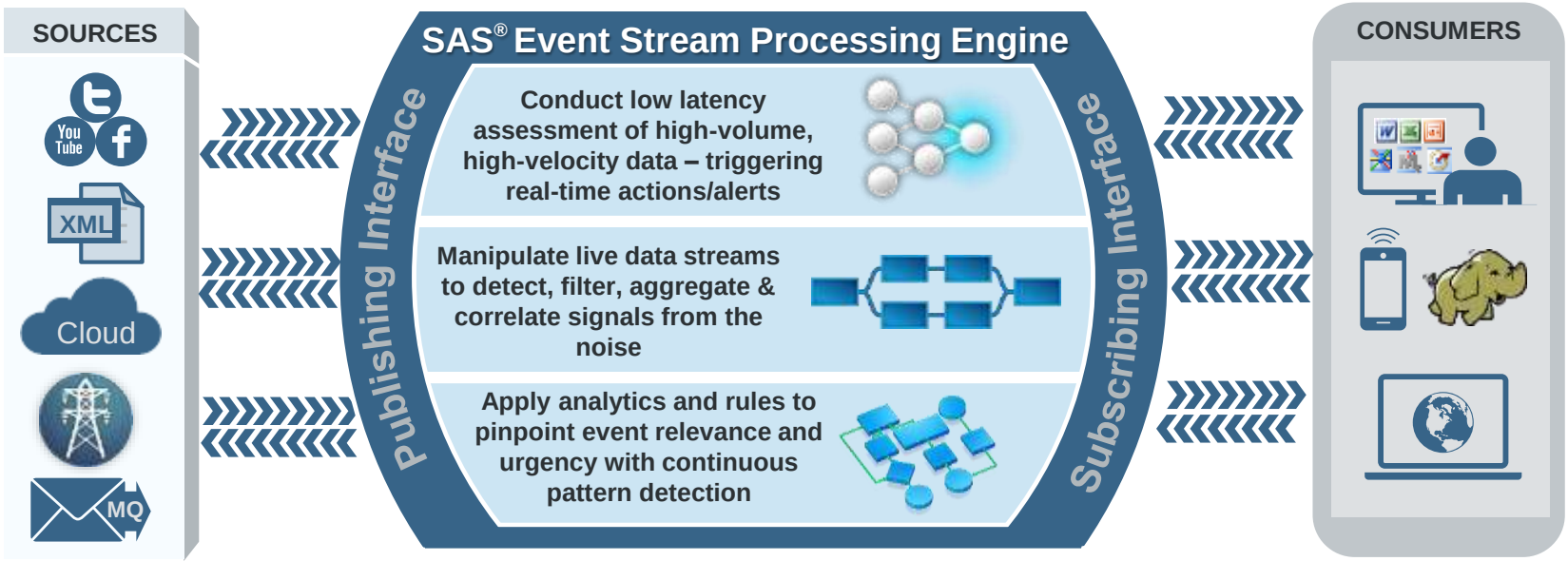
SAS VISUAL ANALYTICS / STATISTICS



50 minutos

PROCESAMIENTO COMPLEJO DE EVENTOS

PROCESAMIENTO "ON-THE-FLY"



PROCESAMIENTO COMPLEJO DE EVENTOS

CASOS DE USO

Optimización E-Commerce

- Análisis de uso de sitios Web
- Optimización de experiencia de usuario
- Publicidad y ofertas en tiempo real

Detección de Fraudes

- Análisis de transacciones
- Alertas en tiempo real y manejo de casos.
- Correlación con comportamiento del cliente.

Internet de las Cosas

- Sensores en tiempo real
- Detección de anomalías en tiempo real.
- Monitoreo de activos críticos
- Instrucciones en tiempo real

Comunicaciones

- Alertas de mensajes en tiempo real.
- Ofertas y acciones en tiempo real.
- Diagnóstico y acciones

Decision Management

- Decisiones operacionales centralizadas inmediatas.
- Directivas a empleados y sistemas en tiempo real.

Mercados de Capital

- Cálculos y monitoreo contínuo.
- Minimización del tiempo entre las operaciones y su reporte.

PASOS PARA EL ARMADO DEL LABORATORIO

1. SAS da soporte al cliente en el diseño del cluster Hadoop y en establecer los criterios de éxito para las pruebas.
2. El cliente pone el cluster Linux (referencia 4 nodos, 8 cores, 128 GB RAM por nodo).
3. El cliente elige la distribución de Hadoop (Hortonworks / Cloudera).
4. SAS entrega el software por el período de prueba (referencia 3 meses):
    1. SAS Data Loader for Hadoop.
    2. SAS Visual Analytics / Statistics.
    3. SAS In-Memory Statistics for Hadoop.
    4. SAS Event Stream Processing.

PASOS PARA EL ARMADO DEL LABORATORIO

5. Servicios profesionales de SAS instalan Hadoop y las soluciones SAS elegidas. Este es el único cargo al cliente a precio promocional.
6. Sugerencia de prueba:
    1. Formatos ORC/Parquet y SAS SPDE.
    2. Hive y SAS Data Loader for Hadoop.

ESPERAMOS QUE NUESTRAS SOLUCIONES SEAN DE SU UTILIDAD