SAS GLOBAL FORUM

Hadoop and SAS

Paul Kent
VP BigData, SAS

Paul.Kent @ sas.com

@hornpolish

paulmkent

SAS and Hadoop :: the BIG Picture

SAS and Hadoop are made for each other

This talk explains some of the reasons why they are a good fit.

Examples are drawn from the customer community to illustrate how SAS is a good addition to your Hadoop Cluster.

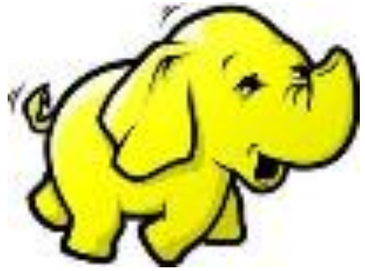
SAS GLOBAL FORUM

The Stages of the Relationship

1. Connecting (Getting to know each other)
   - What exactly is Hadoop?
   - Base SAS connections to Hadoop
2. Dating
   - SAS Access to Hadoop
   - Pig Storage extensions from SAS
3. Committed
   - Data Management Studio for Hadoop
   - SAS High Performance Procedures and the LASR Analytic Server

SAS GLOBAL FORUM

1. Connecting



Getting to know one another…

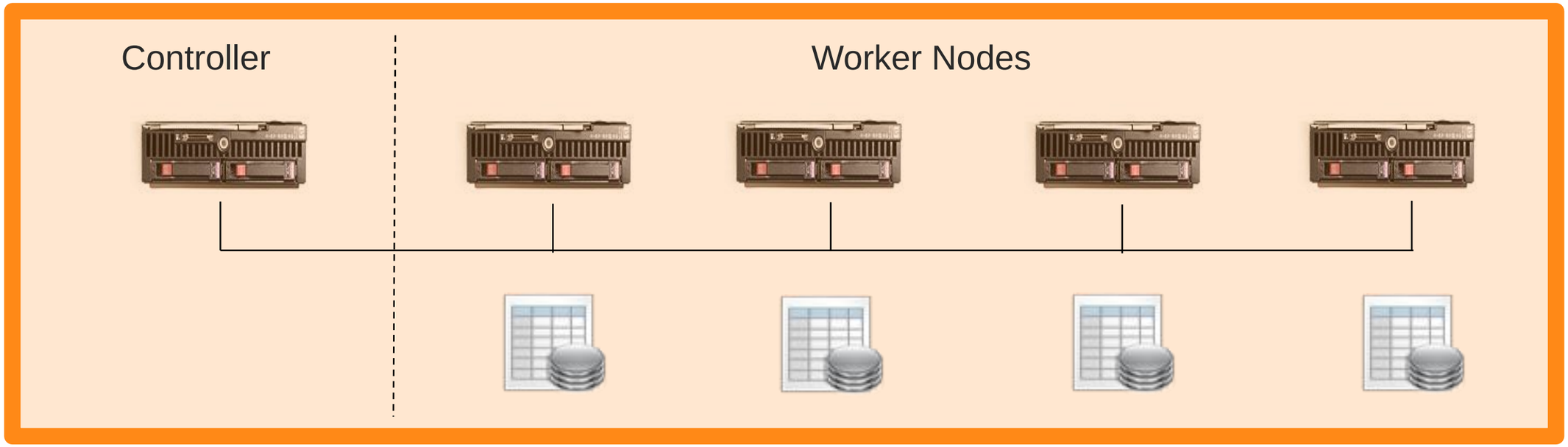
Apache Hadoop

The project includes these subprojects:

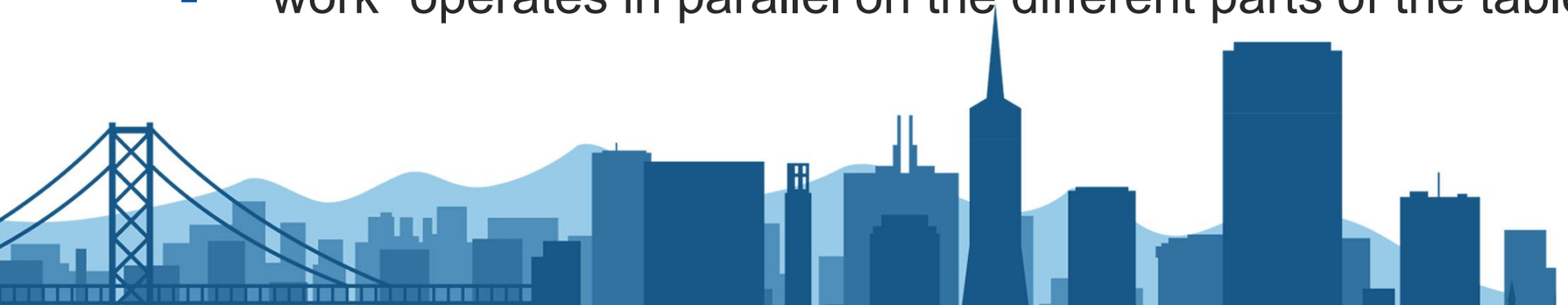
- Hadoop Common: The common utilities that support the other Hadoop subprojects.
- Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
- Hadoop MapReduce: A software framework for distributed processing of large data sets on compute clusters.

SAS GLOBAL FORUM

Hadoop – Simplified View



- MPP (Massively Parallel) hardware running database-like software
- A single logical table is stored in parts across multiple worker nodes
- “work” operates in parallel on the different parts of the table



Idea #1 - HDFS. Never forgets!

Head Node	Data 1	Data 2	Data 3	Data 4…
MYFILE.TXT				
..block1 ->	block1copy1			
..block2 ->		block2copy2		
..block3 ->			block3copy3	

SAS GLOBAL FORUM

Idea #1 - HDFS. Never forgets!

Head Node	Data 1	Data 2	Data 3	Data 4…
MYFILE.TXT				
..block1 ->	block1copy1		block1copy2	
..block2 ->		block2copy2		block2copy2
..block3 ->	block3copy2		block3copy3	

SAS GLOBAL FORUM

Idea #1 - HDFS. Never forgets!

Head Node	Data 1	Data 2	Data 3	Data 4…
MYFILE.TXT				
..block1 ->	block1copy1		bloc cop 2	
..block2 ->		block2copy2		block2copy2
..block3 ->	block3copy2		blo cop 3	

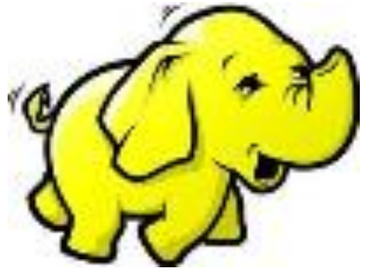
SAS GLOBAL FORUM

Idea #2 - MapReduce

- We Want the Minimum Age in the Room

- Each Row in the audience is a data node
- I'll be the coordinator
  - From outside to center, accumulate MIN
  - Sweep from back to front. Youngest Advances

1. Connecting



sas

Making a Connection

SAS GLOBAL FORUM

FILENAME xxx HADOOP

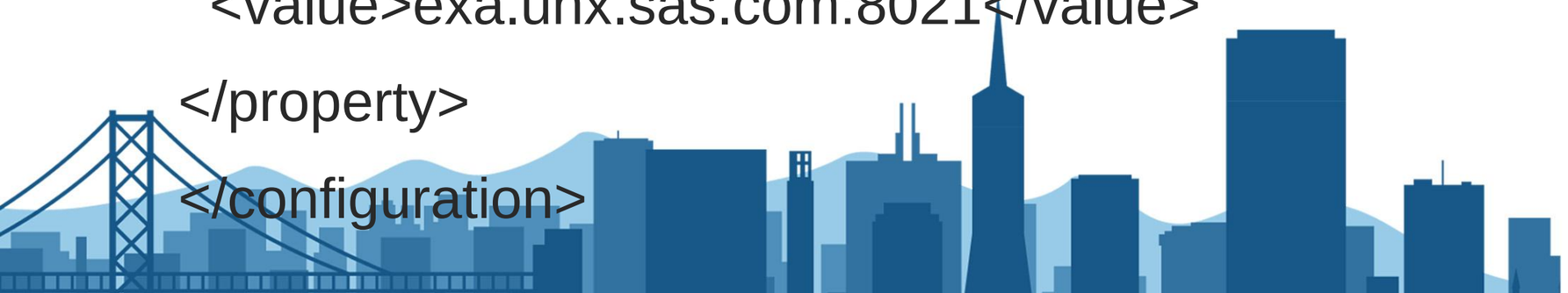
```
FILENAME paul HADOOP
    “/users/kent/mybigfile.txt”
    CONFIG=“/etc/hadoop.cfg” USER=“kent” PASS=“sekrit”;


DATA MYFILE;
    INFILE paul;
    INPUT name $ age sex $ height weight;
    RUN;
```

SAS GLOBAL FORUM

/etc/hadoop.cfg ?

```
<configuration>
<property>
  <name>fs.default.name</name>
  <value>hdfs://exa.unx.sas.com:8020</value>
</property>
<property>
  <name>mapred.job.tracker</name>
  <value>exa.unx.sas.com:8021</value>
</property>
</configuration>
```



Different Hadoop Versions?

```
options set=SAS_HADOOP_JAR_PATH="/u/kent/jars/cdh4/";
```

- OpenSource Apache
- Cloudera CDH3 and CDH4
- Pivotal HD (was Greenplum)
- MAPR
- Hortonworks (including DDN and Teradata OEM editions)
- Intel

SAS GLOBAL FORUM

2. Dating

SAS Learns Hadoop Tables

Hadoop Learns SAS Tables

SAS GLOBAL FORUM

LIBNAME xxx HADOOP

```
LIBNAME olly HADOOP
    SERVER=olly.mycompany.com
    USER="kent" PASS="sekrit";

PROC DATASETS LIB=OLLY;
    RUN;
```

LIBNAME xxx HADOOP

- Cool! I don’t have to repeat the INPUT statement in every program that I want to access my files!!
- Thanks to Apache HIVE
  - supplies the metadata that projects a relational view of several underlying file types.
  - Provides SQL with relational primitives like JOIN and GROUP BY

Hadoop LIBNAME Statement

SAS Server

Hadoop Cluster

```
LIBNANE olly HADOOP
  SERVER=hadoop.company.com
  USER="paul" PASS="sekrit"

PROC MEANS DATA=olly.table;
  RUN;
```

Select *
From olly

Hadoop
Access
Method

Controller

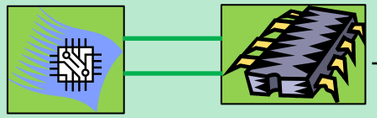
Workers

Select *
From olly

Select *
From olly_slice

Potentially

Big Data



SAS GLOBAL FORUM

Hadoop LIBNAME Statement – with SQL Pasthru

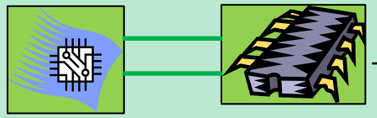
SAS Server

Hadoop Cluster

```
LIBNANE olly HADOOP
  SERVER=hadoop.company.com
  USER="paul" PASS="sekrit"

PROC MEANS DATA=olly.table;
  RUN;
```

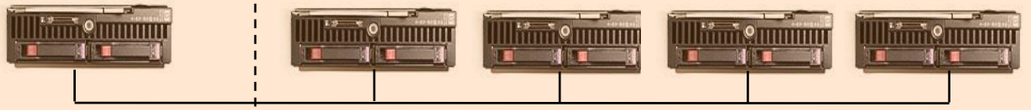
Select sum(x),
min(x) ….
From olly



Hadoop
Access
Method

Controller

Workers



Select sum(x),
min(x) …
From olly

Select sum(x),
min(x) ….
From olly_slice

Aggregate Data
ONLY

SAS GLOBAL FORUM

HADOOP LIBNAME Statement

- PROC SQL explicit SQL is supported
- This sends the SQL exactly as you typed it down into the HIVE processor
- One way to move the work (joins, group by) down onto the cluster

SAS GLOBAL FORUM

Hadoop (PIG) Learns SAS Tables

```
register pigudf.jar, sas.lasr.hadoop.jar, sas.lasr.jar;
/* Load the data from a CSV in HDFS */
A = load '/user/kent/class.csv'
    using PigStorage(',')
    as (name:chararray, sex:chararray,
        age:int, height:double, weight:double);
```

(continued…)

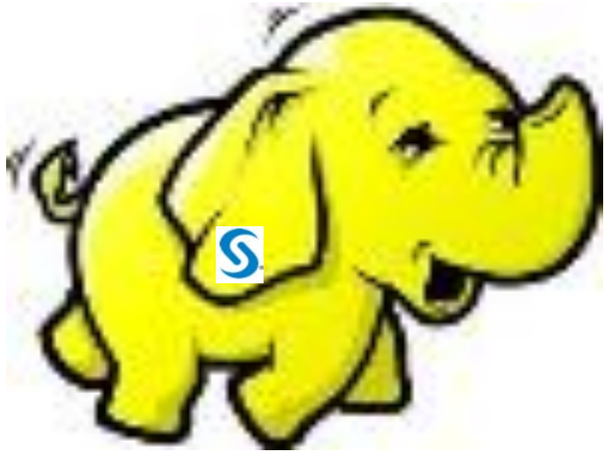
SAS GLOBAL FORUM

Hadoop (PIG) Learns SAS Tables

```
Store A into '/user/kent/class'
        using com.sas.pigudf.sashdat.pig.SASHdatStoreFunc(
            ’bigcdh01.unx.sas.com',
            '/user/kent/class_bigcdh01.xml');
```

SAS GLOBAL FORUM

3. Committed

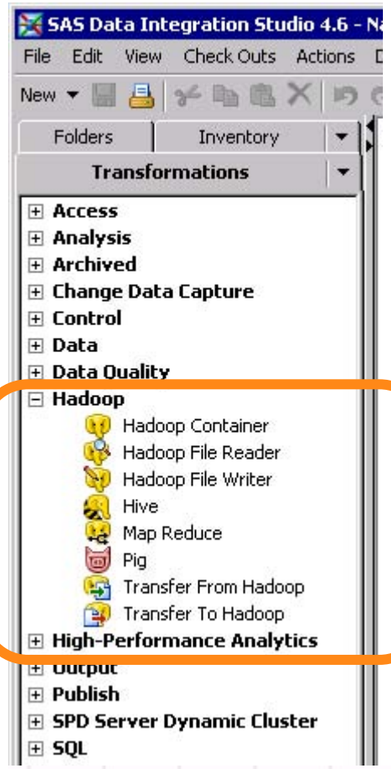


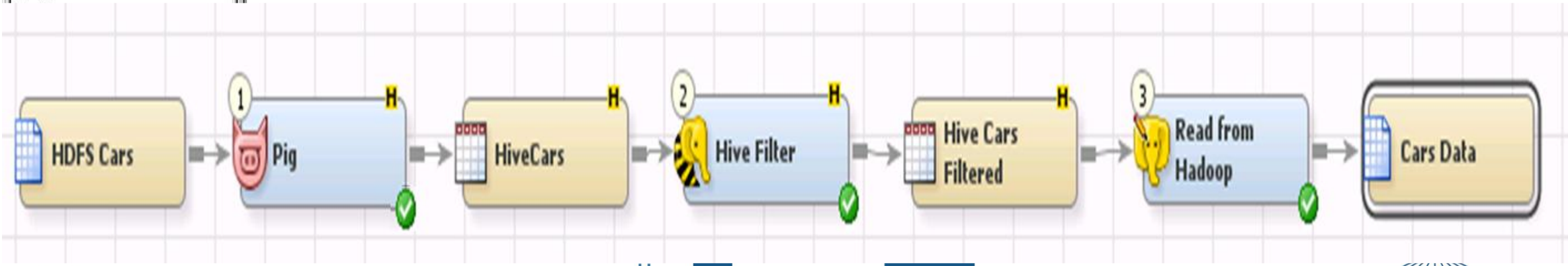
Data Management for Hadoop

SAS HPA and VA on Hadoop

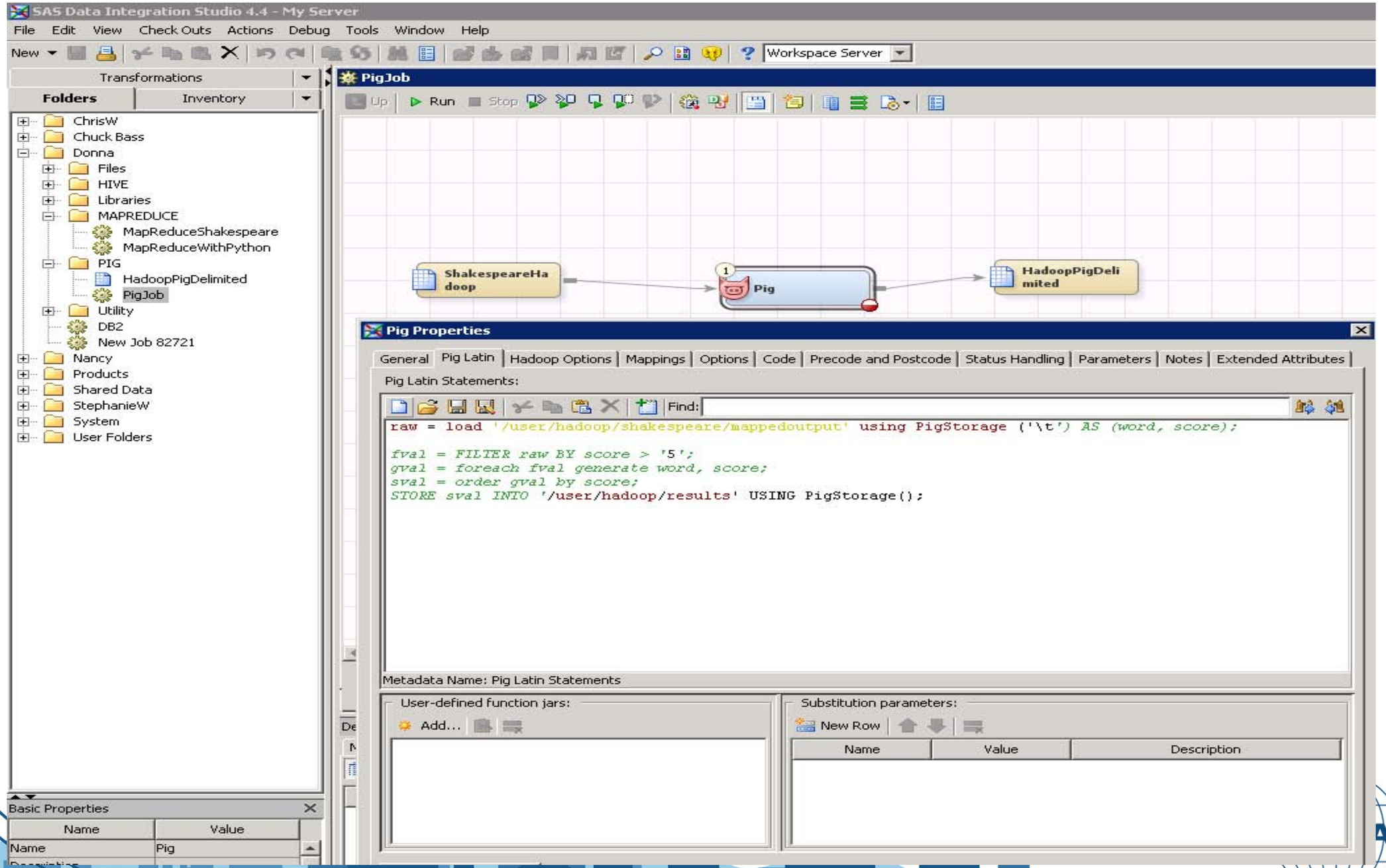
SAS GLOBAL FORUM

Hadoop transforms in DI Studio

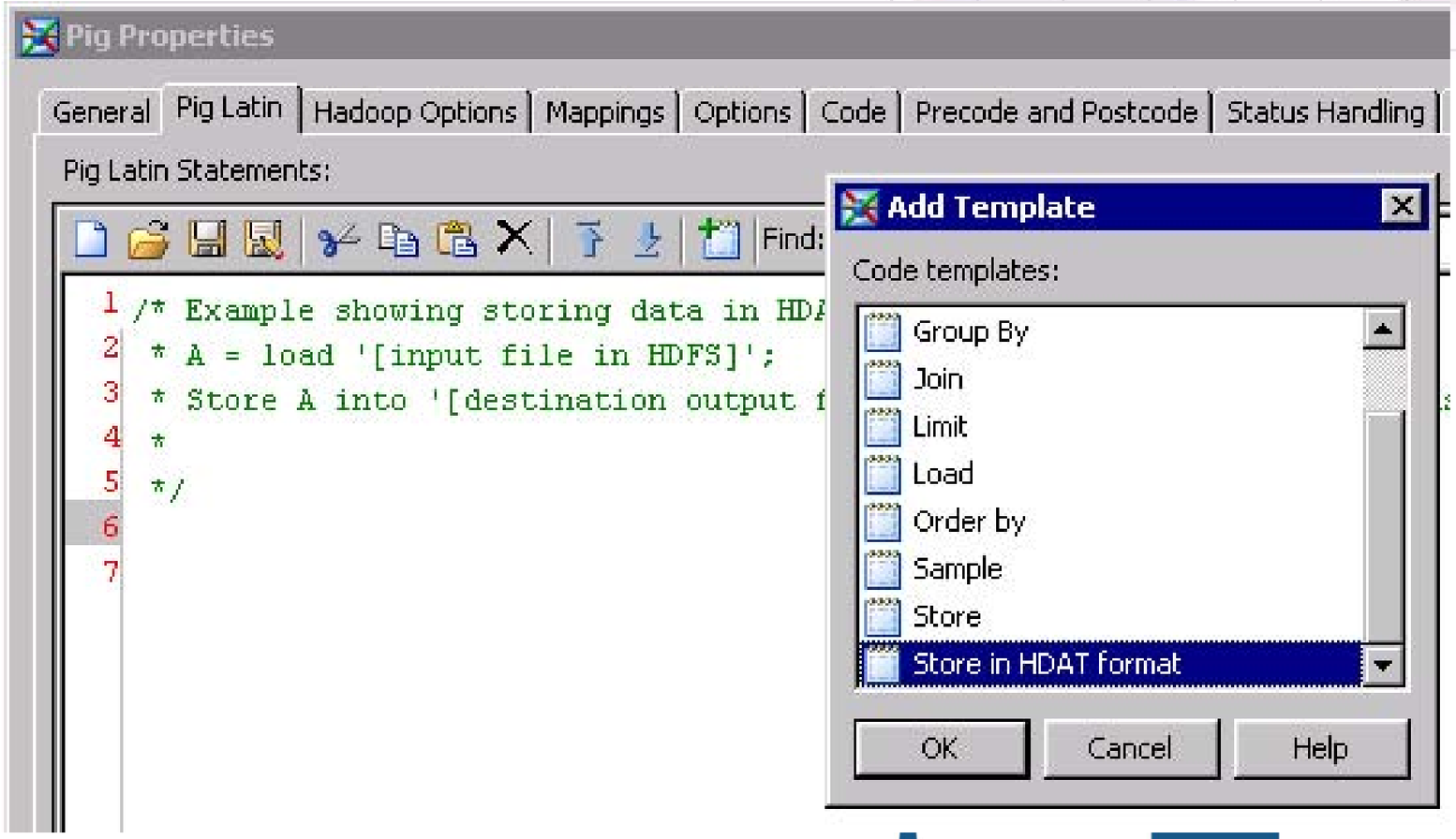




SAS GLOBAL FORUM


SAS Data Integration Studio 4.4 - My Server
File Edit View Check Outs Actions Debug Tools Window Help
New
Workspace Server
Transformations
Folders
Inventory
ChrisW
Chuck Bass
Donna
Files
HIVE
Libraries
MAPREDUCE
MapReduceShakespeare
MapReduceWithPython
PIG
HadoopPigDelimited
PigJob
Utility
DB2
New Job 82721
Nancy
Products
Shared Data
StephanieW
System
User Folders
Basic Properties
Name
Value
Name
Pig
PigJob
Up
Run
Stop
ShakespeareHadoop
1
Pig
HadoopPigDelimited
Pig Properties
General
Pig Latin
Hadoop Options
Mappings
Options
Code
Precode and Postcode
Status Handling
Parameters
Notes
Extended Attributes
Pig Latin Statements:
Find:
raw = load '/user/hadoop/shakespeare/mappedoutput' using PigStorage ('\t') AS (word, score);
fval = FILTER raw BY score > '5';
gval = foreach fval generate word, score;
sval = order gval by score;
STORE sval INTO '/user/hadoop/results' USING PigStorage();
Metadata Name: Pig Latin Statements
User-defined function jars:
Add...
Substitution parameters:
New Row
Name
Value
Description

Data Integration and LASR



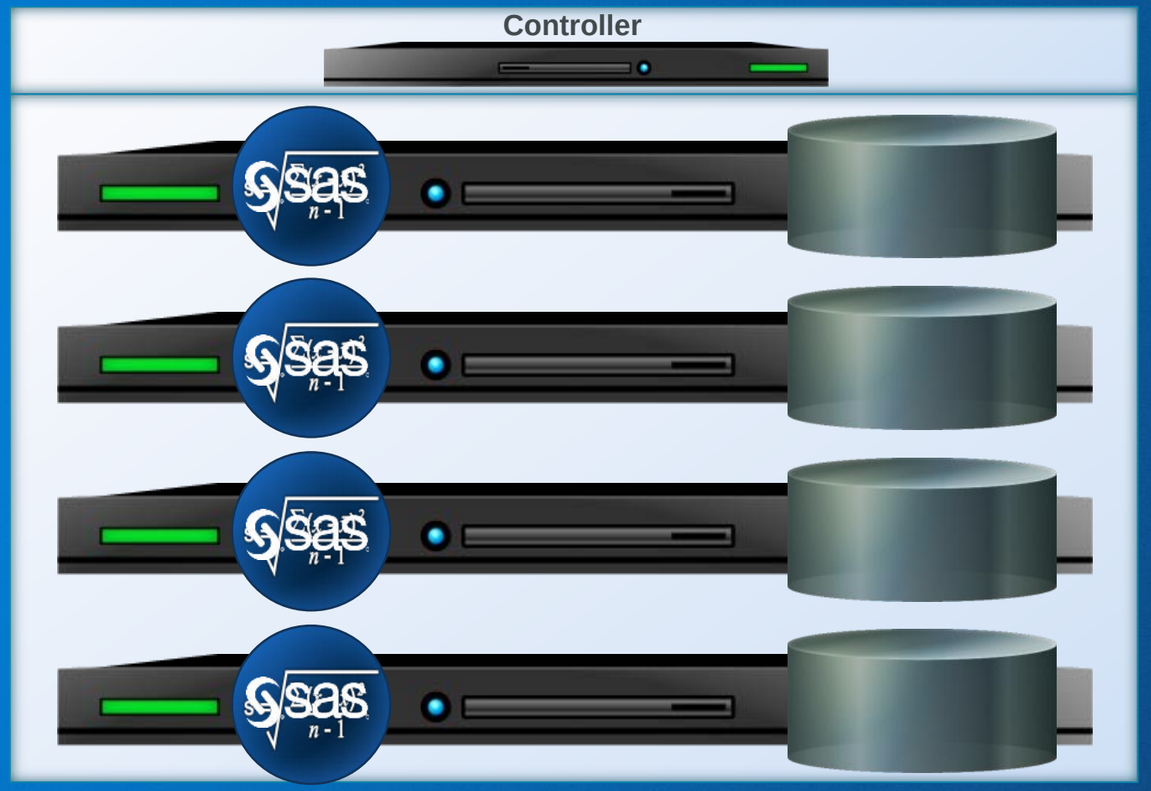
SAS GLOBAL FORUM

SAS HPA AND VA ON HADOOP

Client

HPA ALONGSIDE

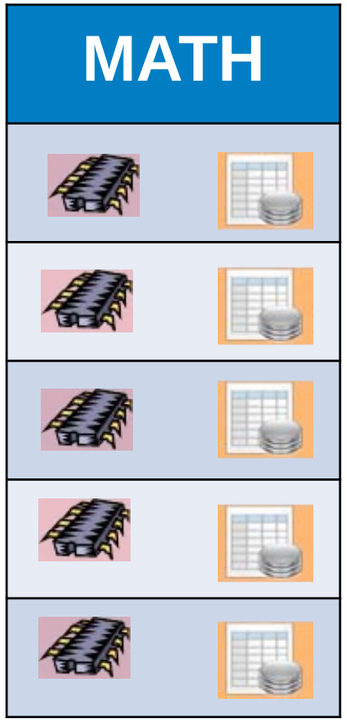
Client



DATA INSIDE APPLIANCE

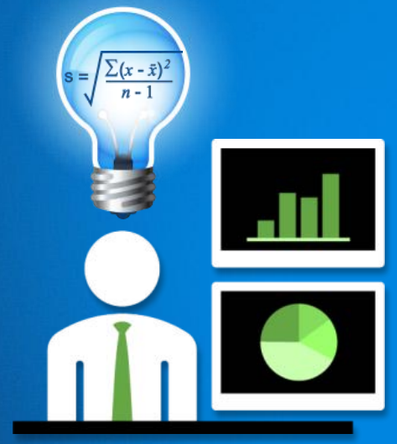
SAS Client

Visual Analytics

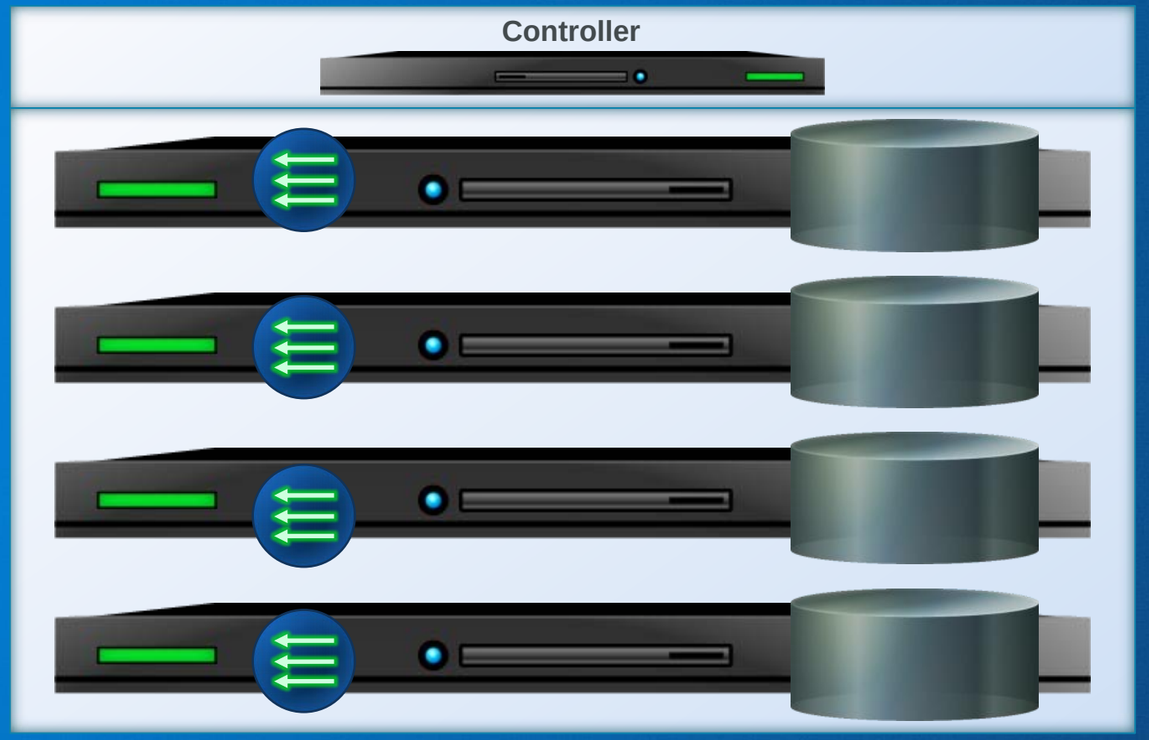


2012

HPA ASYMMETRIC



Client



DATA FROM ANYWHERE

SAS Client

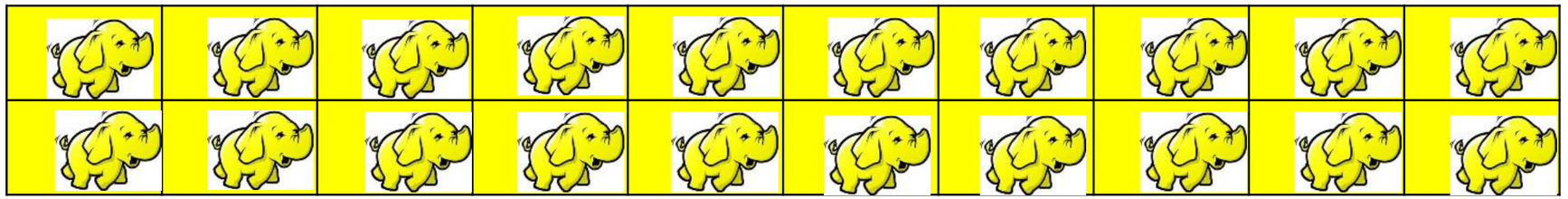
Visual Analytics

MATH

EDW

EDW

2013+

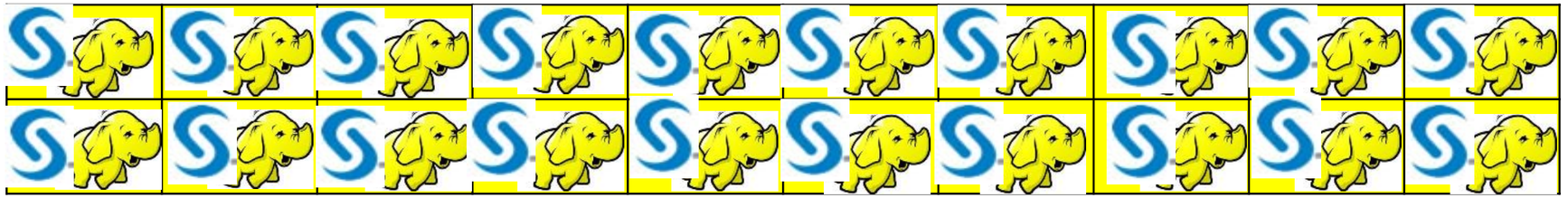


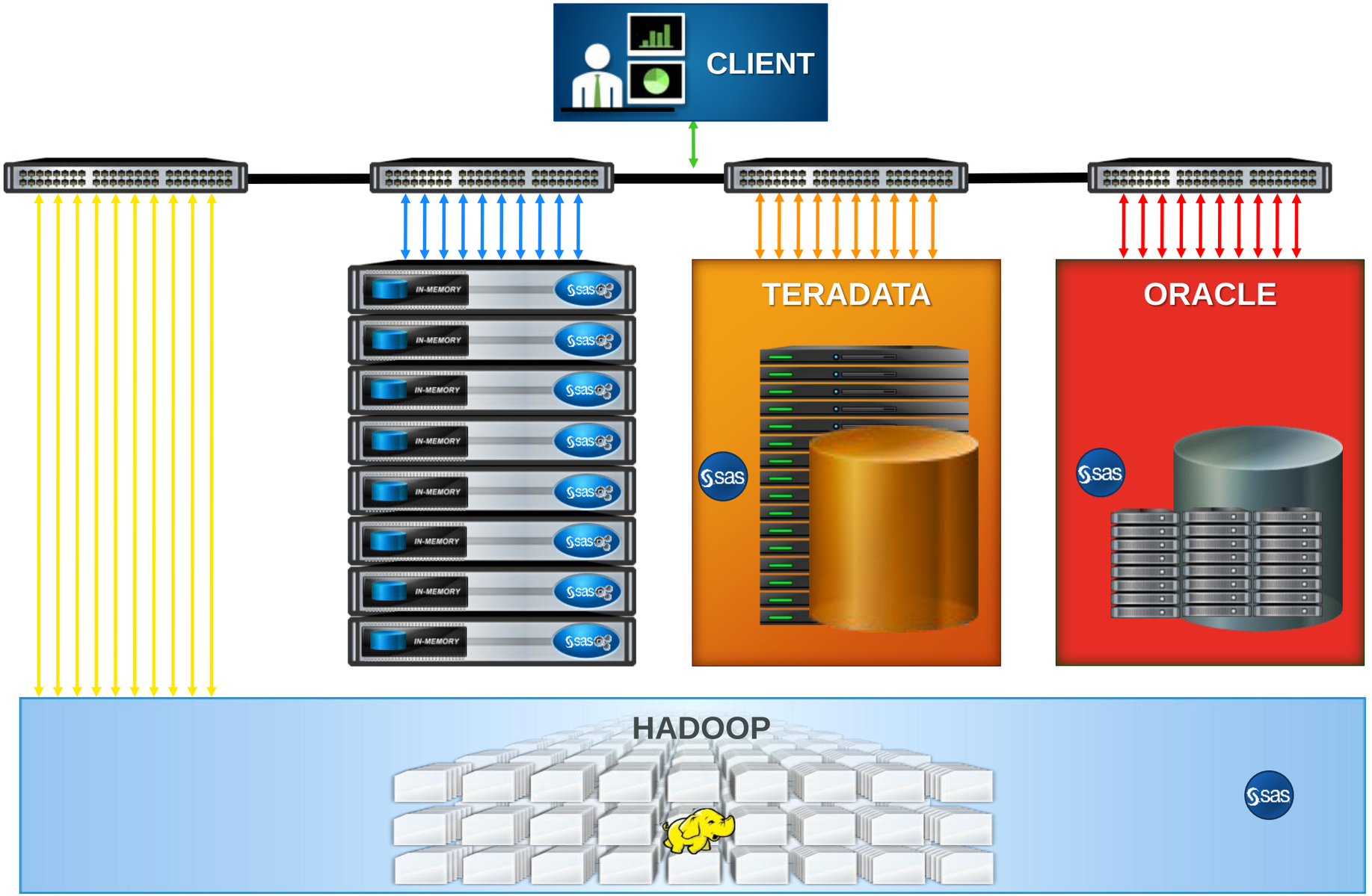
DATA FROM ANYWHERE (II)

SAS Client

Visual Analytics

2013+





BIG DATA • What's in it for me?

1. Increase cycle time on your existing datasets
2. Use your existing data in more complex ways
3. Capture and Process new datastreams
4. Use ALL of your data

HIGH PERFORMANCE VISUALIZATION

Scan rate:

- 1 billion records per second

Analytics:

- Summarization of 1 billion records 0.2 seconds
- 45 simultaneous pairs of correlations on 1 billion records in ~ 5 seconds

“Billion is the new million”

Paul.Kent @ sas.com

@hornpolish

paulmkent

Thank You!

SAS® GLOBAL FORUM

San Francisco, CA

April 28–May 1, 2013