SAS GLOBAL FORUM

SAS and Hadoop

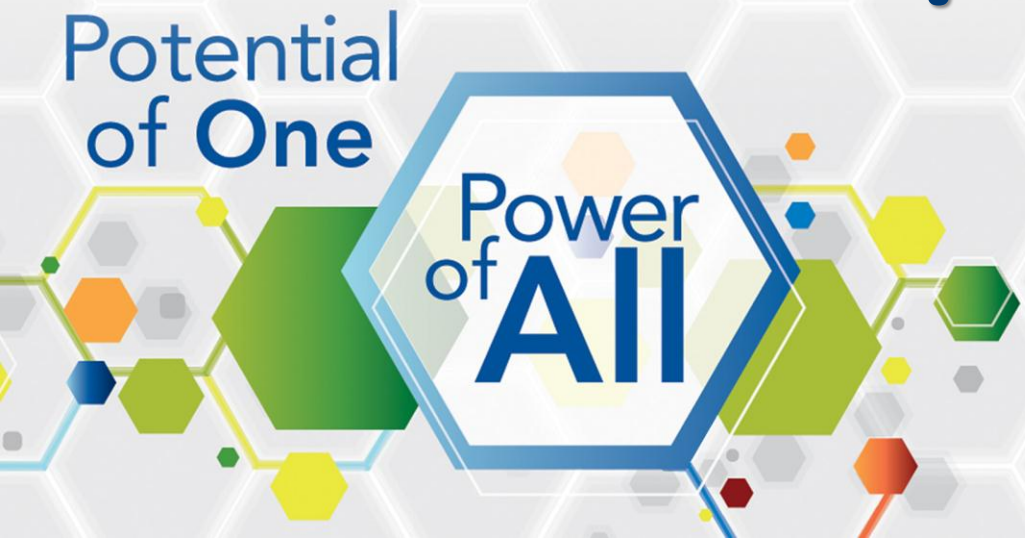
3rd Annual State of the Union

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.

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. Engaged
   - Scoring Accelerator
4. Committed
   - Data Management Studio for Hadoop
   - SAS High Performance Procedures and the LASR Analytic Server

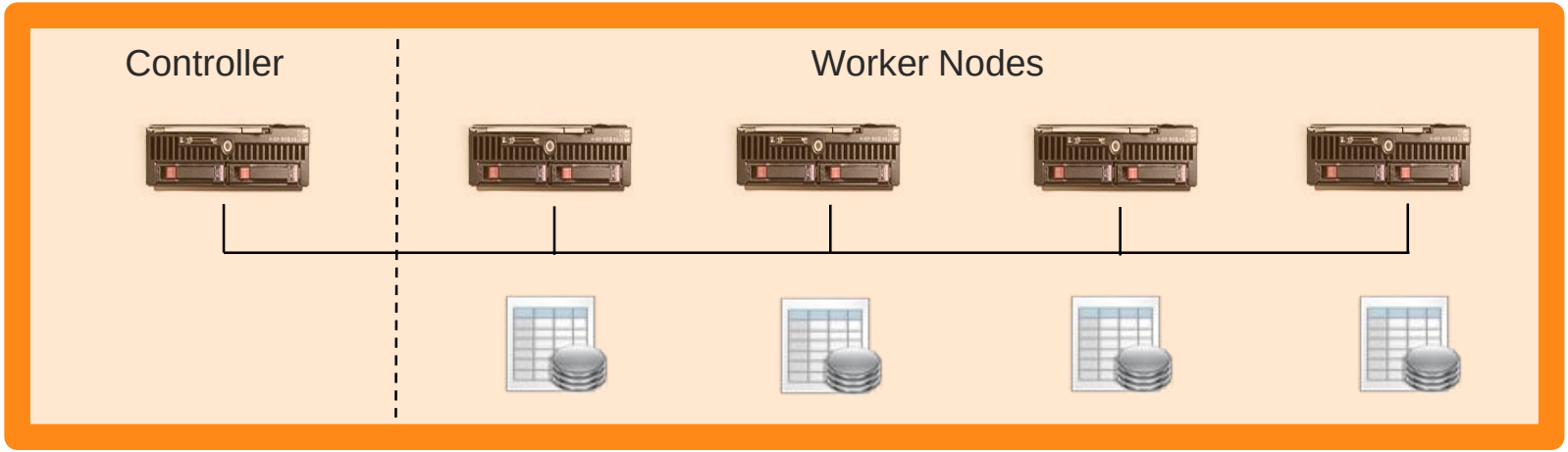
1. Introductions

Apache Hadoop - Background

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.

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	

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	

Idea #1 - HDFS. Never forgets!

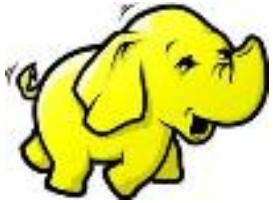
Head Node	Data 1	Data 2	Data 3	Data 4…
MYFILE.TXT				
..block1 ->	block1copy1		blo[illegible]2	
..block2 ->		block2copy2		block2copy2
..block3 ->	block3copy2		blo[illegible]3	

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



Making a Connection

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;
```

/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?

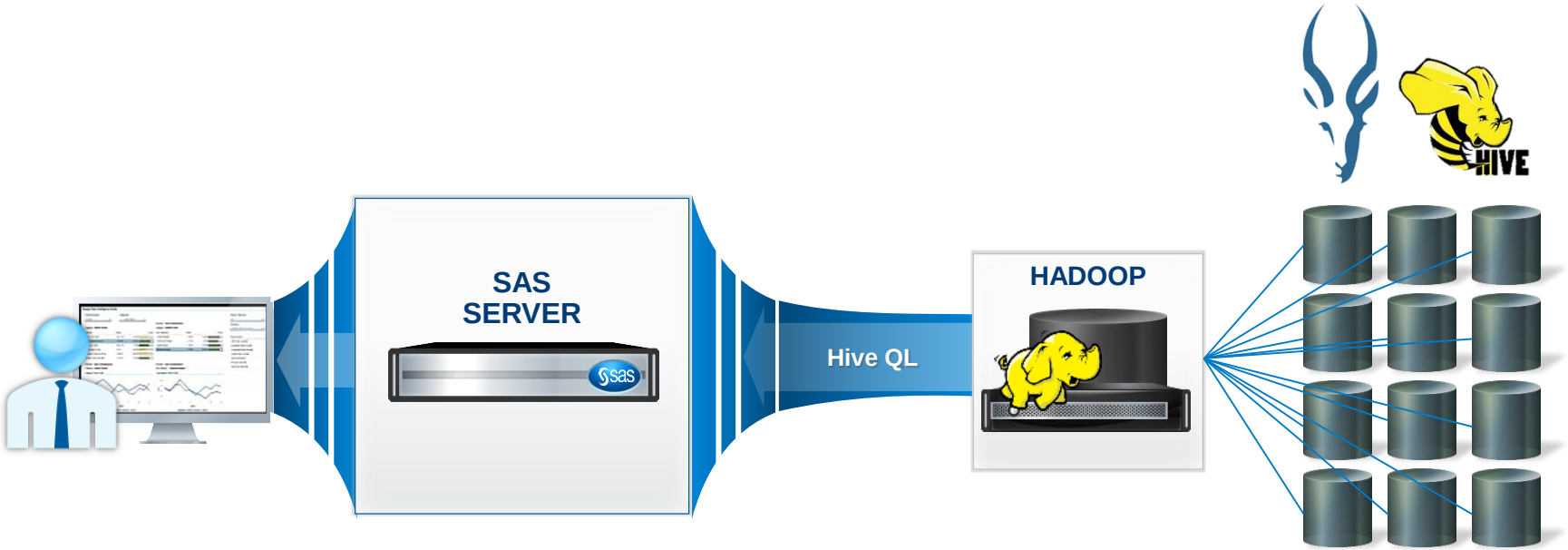
options set=SAS_HADOOP_JAR_PATH="/u/kent/jars/cdh4/”;

- OpenSource Apache 2.0
- Cloudera CDH4 and CDH5
- Hortonworks 1.3.2 and 2.x (including DDN and Teradata OEM editions)
- Pivotal HD (was Greenplum)
- Intel
- MAPR

2. Dating

SAS Learns Hadoop Tables

Hadoop Learns SAS Tables


SAS
SERVER
Hive QL
HADOOP
HIVE

LIBNAME xxx HADOOP

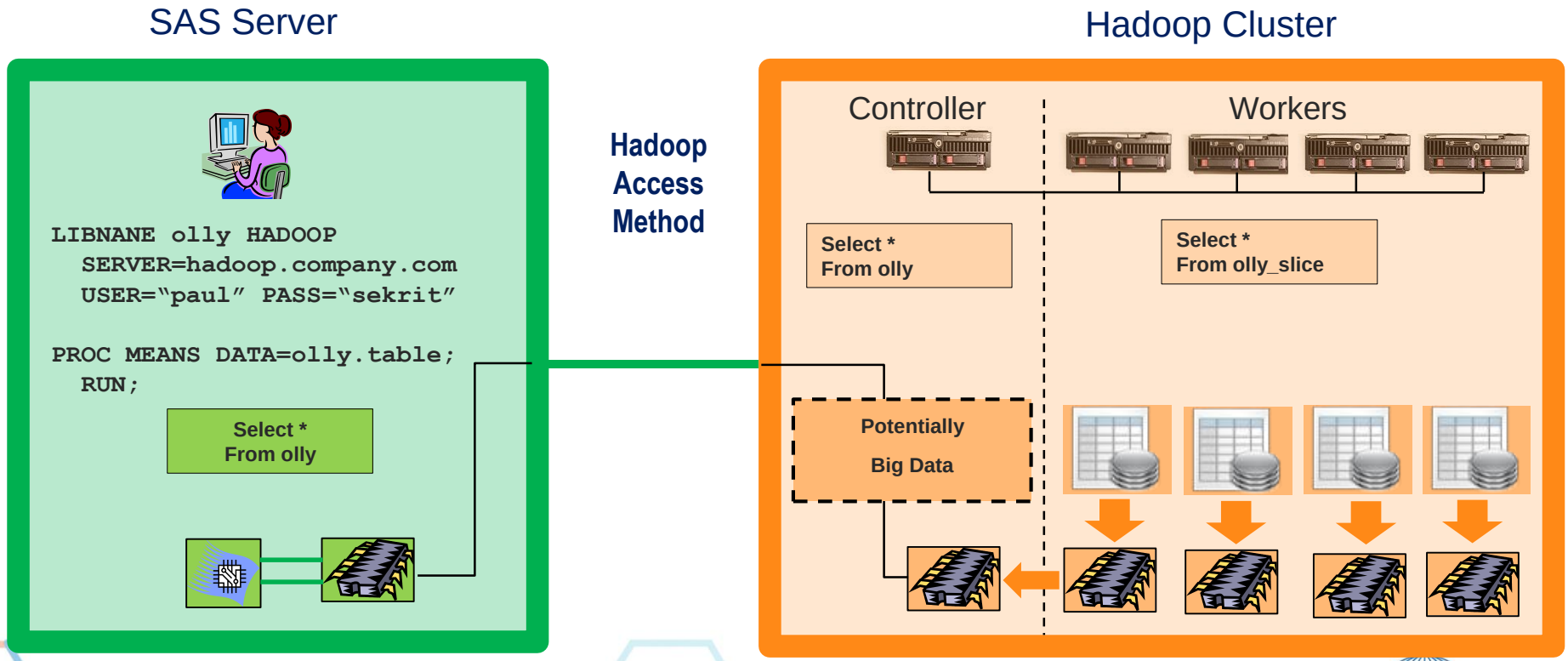
```
LIBNAME olly HADOOP
    SERVER=olly.mycompany.com
    USER="kent" PASS="sekrit";

PROC DATASETS LIB=OLLY;
    RUN;
```

LIBNAME xxx HADOOP

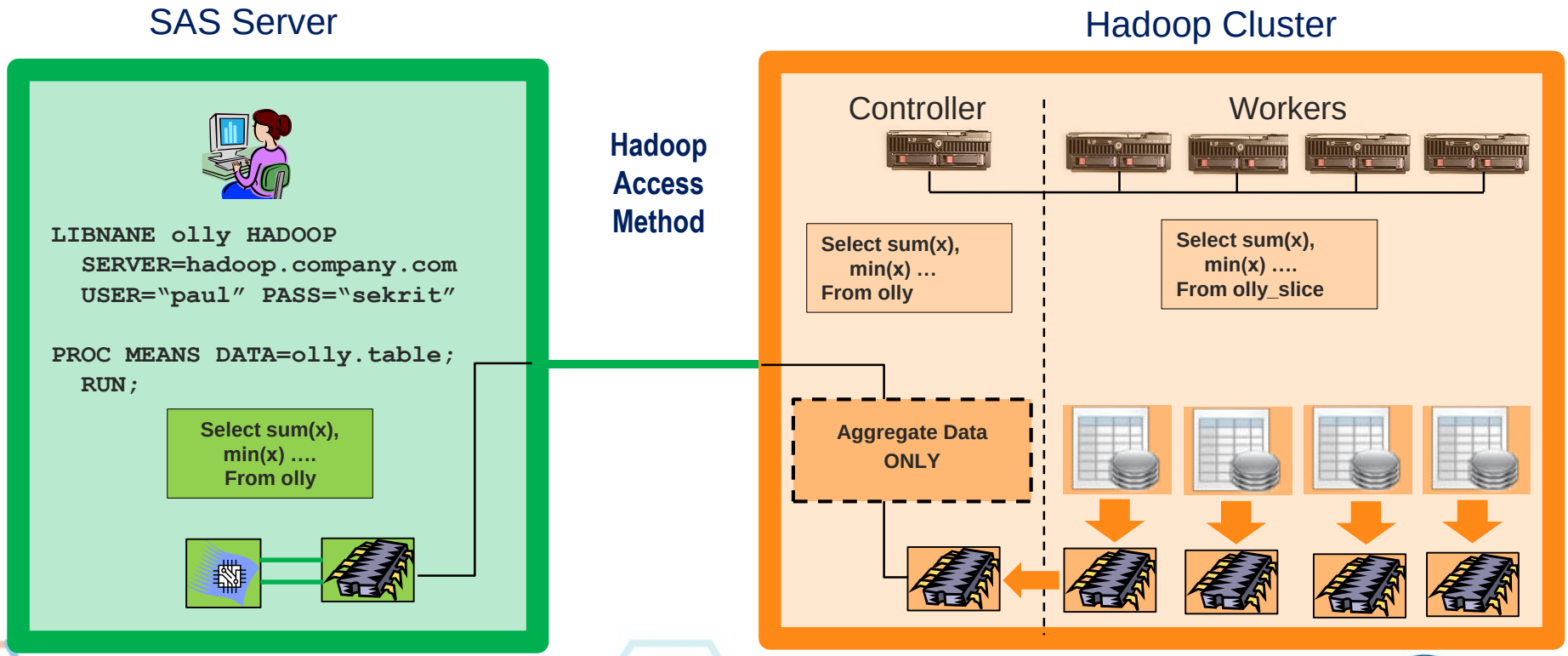
- Cool! I don't have to repeat the INPUT statement in every program that I want to access my files!!
- Thanks to Apache HIVE
  - supplies the metadata that projects a relational view of several underlying file types.
  - Provides SQL with relational primitives like JOIN and GROUP BY

Not only HIVE. Cloudera Impala

Hadoop LIBNAME Statement



Hadoop LIBNAME Statement – with SQL Pasthru



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

HADOOP LIBNAME Statement

- PROC SQL implicit SQL is supported
- Base Procedure pushdown
- More ways to move the work down onto the cluster

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…)

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');
```

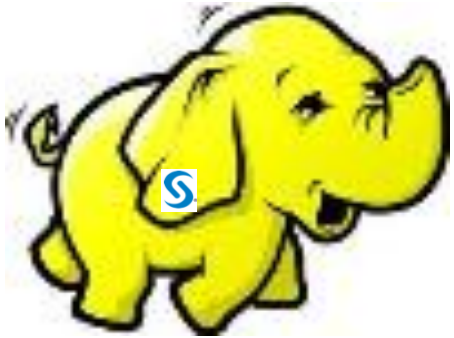
SAS Hadoop Engine – Anyfile Reader

- Consistent with Hadoop and its laissez-faire approach to schema “on Need”
- PROC HDMD describes as much of the file as you want to see as a table
- Libname engine matches metadata (HDMD) with datafile and returns “rowbuffers” to calling SAS procedure
- Imagine exporting INFILE/INPUT statement to Hadoop to be run there (at hadoop scale)

TEACH HADOOP (MAP/REDUCE) ABOUT SAS

```
/* Create HDMD file */
proc hdmd name=gridlib.people
        format=delimited
        sep=tab
        file_type=custom_sequence
        input_format='com.sas.hadoop.ep.inputformat.sequence.PeopleCustomSequenceInputFormat'
        data_file='people.seq';

    COLUMN name   varchar(20) ctype=char;
    COLUMN sex    varchar(1)  ctype=char;
    COLUMN age    int         ctype=int32;
    column height double      ctype=double;
    column weight double      ctype=double;
run;
```

3. Engaged



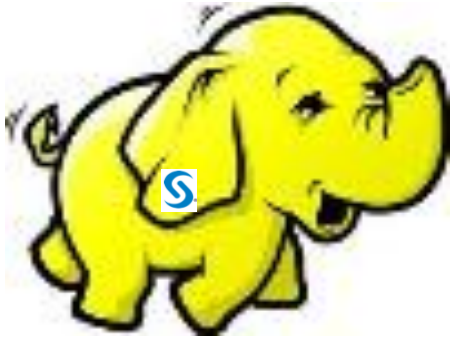
SAS Embedded Process

SAS / Embedded Process

SAS/Scoring Accelerator for Hadoop
SAS/Code Accelerator for Hadoop (July 7th 2014)
SAS/Data Quality Accelerator for Hadoop (July 7th 2014)
SAS/Data Director* (Name TBD July 7th 2014)

```
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;
```

4. Committed



Data Management for Hadoop

SAS HPA and VA on Hadoop

Solutions on Hadoop

SAS DATA DIRECTOR
JUNE' 2014

- ETL in Hadoop via SAS
  - Parallel Data Step
  - Read / Write Text & HDFS Files
  - Text Processing in Hadoop
  - SQOOP Integration
  - Structure and Prepare Data For Analysis
- Play nicely in the Hadoop Ecosystem
  - MR1 & MR2 support
- Play nicely with SAS
  - Libname, Procs & SAS Code
- Design Patterns
  - Input / output Patterns
  - Filtering, Expressions & Summarization
  - Data Structuring & Organizing
    - » Sort, Join, Merge, Pivot/Transpose
- Job Chaining & Streaming
- Data Quality
  - Parse, Standardize, Match etc…
  - Data Profiling

What directive do you want to perform?

Show: **All Directives**

Saved Directives
Open a previously created directive to run, view, or edit.

Schedule a Directive to Run
Schedule a directive to run at specified dates and times

Chain Directives Together
Run a number of directives in a specific order.

Copy Data for Visualization
Copy data from Hadoop and load it into LASR for visualization. Existing data in the target table will be replaced.

Copy Data to Hadoop
Copy data from a source and load it into Hadoop. Existing data in the target file will be replaced.

Join Tables in Hadoop
Create a table in Hadoop from multiple tables.

Pivot a Table in Hadoop
Transpose the columns of a table in Hadoop.

Transform Data in Hadoop
Transform the data in an Hadoop data file.

1 Click

Verify Mailing Address
Check the validity of the mailing address data in a table.

Profile Data
Create a report profiling the data in a table.

Generate Business Rules
Analyze data in a table and generate business rules.

Send Data for Remediation
Select data to send to the remediation queue for further action.

Transform Data in Hadoop

Back to Directives | Save | Save As…

Source Table

Choose the data source with data you want to transform

New Data Source… | Manage Data Sources…

Customer Information
Master data for all customer information

Sales Associates
Contains sales data for all the sales associates in the company

Sales Data
Contains historic sales results for financial planning

Market Research
Market analysis data for all divisions in the company

Mock Data
This data has been created for test purposes only

2012 Sales
Create a table in Hadoop from multiple tables.

Next

Transform Data in Hadoop

Back to Directives | Save | Save As…

Source Table: Sales Data

Select the table with the data you want to transform

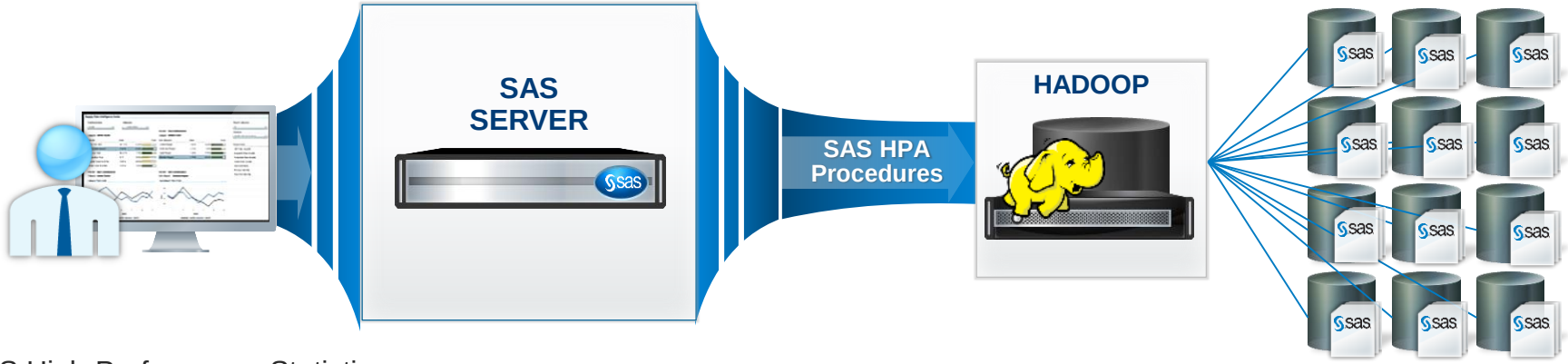
Back to Data Sources | Filter…

Argentina	Australia	Bangladesh	Belgium	Brazil
Bulgaria	Canada	Chile	Croatia	Denmark

1 Click

Next

SAS / High Performance Analytics



SAS High-Performance Statistics
SAS High-Performance Data Mining
SAS High-Performance Text Mining
SAS High-Performance Econometrics
SAS High-Performance Forecasting
SAS High-Performance Optimization

SAS / High Performance Analytics

Prepare

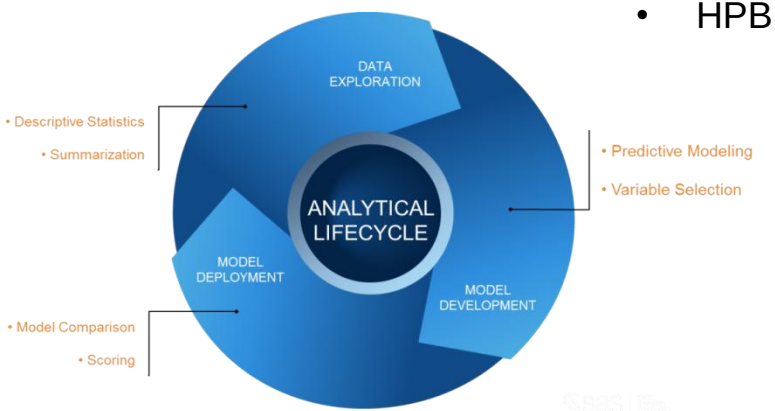
- HPDS2
- HPDMDB
- HPSAMPLE

Explore / Transform

- HPSUMMARY
- HPCORR
- HPREDUCE
- HPIMPUTE
- HPBIN

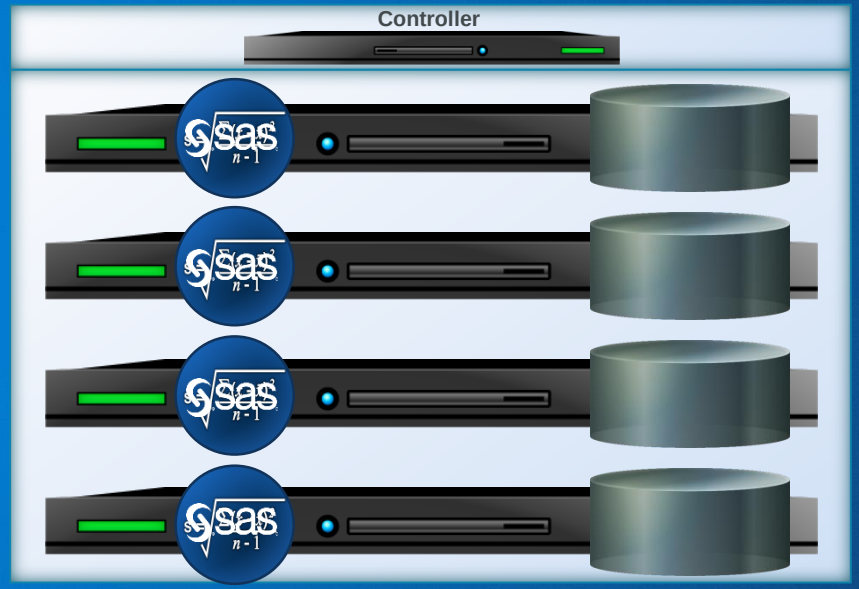
Model

- HPLOGISTIC
- HPREG
- HPNEURAL
- HPNLIN
- HPCOUNTREG
- HPMIXED
- HPSEVERITY
- HPFOREST
- HPSVM
- HPDECIDE
- HPQLIM
- HPLSO
- HPSPLIT
- HPTMINE
- HPTMSCORE



SAS / High Performance Analytics

Client

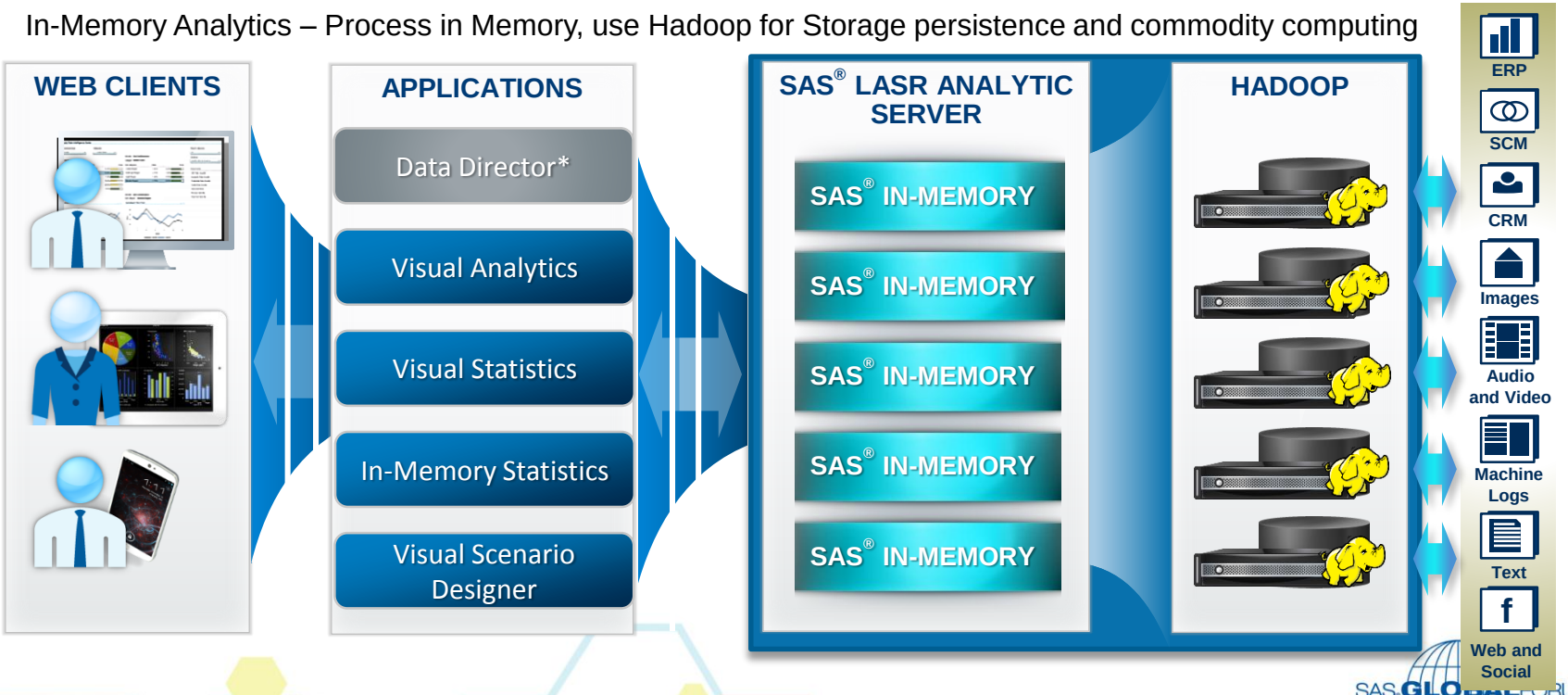


IN-MEMORY(LASR BASED) SOLUTIONS ON HADOOP

4

SAS ANALYTIC HADOOP ENVIRONMENT

In-Memory Analytics – Process in Memory, use Hadoop for Storage persistence and commodity computing



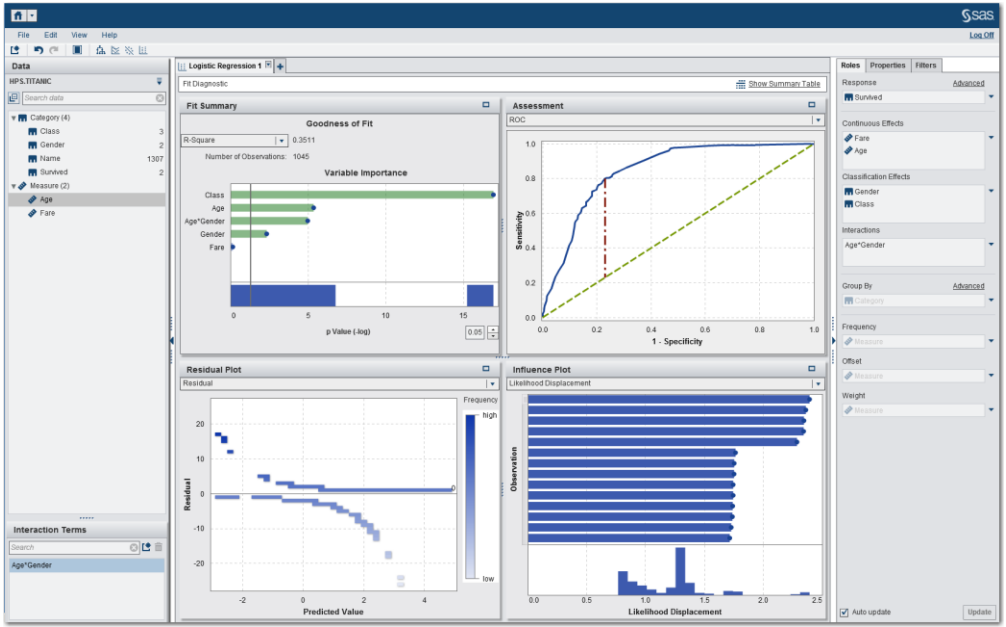
SAS VISUAL ANALYTICS

- Interactive exploration, dashboards and reporting
- Auto-charting automatically picks the best graph
- Forecasting, scenario analysis, Decision Trees and other analytic visualizations
- Text analysis and content categorization
- Feature-rich mobile apps for iPad® and Android

SAS VISUAL STATISTICS

JUNE-2014

new and improved

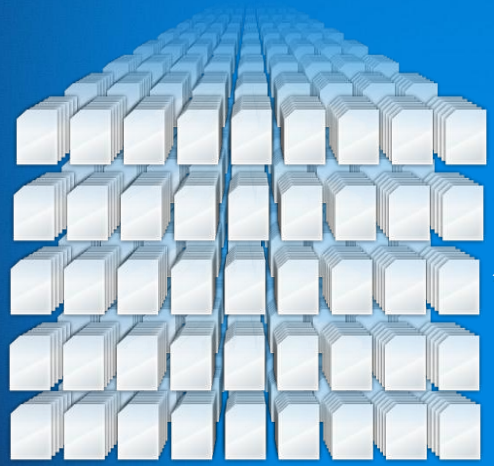


- Interactive, visual application for statistical modeling and classification
- Multiple methods:
  - logistic, Regression, GLM, Trees, Forest, Clustering and more…
- Model comparison and assessment
- Group BY Processing

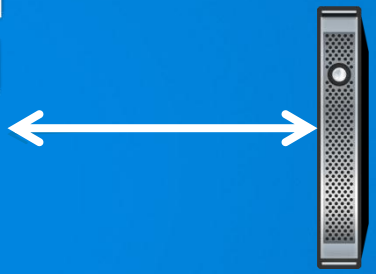
SAS In-Memory Statistics (IMSTAT) for Hadoop March'14

- Data Prep in Hadoop
  - Parallel Data Step
  - Read / Write Text & HDFS Files
  - Text Processing in Hadoop
  - Structure and Prepare Data For Analysis
- Play nicely in the Hadoop Ecosystem
- Play nicely with SAS Users
  - Libname, Procs & SAS Code
- In-Memory Analytics
  - Interactive "Modern" Analytic Methods
    - » Descriptive Statistics
    - » Statistics
    - » Forecasting
    - » Classification
    - » Text Mining
    - » Optimization
    - » Collaborative Filtering
  - Group By Processing

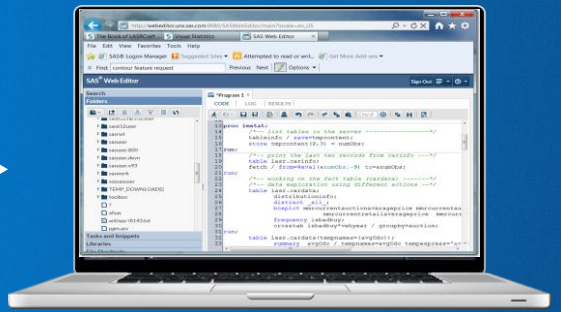
INTERACTIVE ANALYSIS AT SCALE …



LASR Analytic Server on CDH Hadoop



SAS Server
~ BASE, ODS, STAT, Access, LASR
(IMSTAT, RECOMMEND Etc..)



SAS Studio, HTML 5 new Modern Coding Environment

SAS In-Memory Statistics for HAdoop

```
proc imstat;
    /*-- list tables in the server -----------------*/
    tableinfo / save=tmpcontent;
    store tmpcontent(2,3) = numObs;
run;
    /*-- print the last ten records from carinfo ---*/
    table lasr.carinfo;
    fetch / from=%eval(&numObs.-9) to=&numObs;
run;
    /*-- working on the fact table (cardata) -------*/
    /*-- data exploration using different actions --*/
    table lasr.cardata;
         distributioninfo;
         distinct _all_;
         boxplot mmrcurrentauctionaverageprice
            mmrcurrentauctioncleanprice
            mmrcurrentretailaverageprice
            mmrcurrentretailcleanprice /
            groupby=auction;
         frequency isbadbuy;
         crosstab isbadbuy*vehyear / groupby=auction;
run;
    table lasr.cardata(tempnames=(avgOdo));
         summary  avgOdo / tempnames=avgOdo
            tempexpress="avgOdo = vehodo /
            (year(purchdate)-vehyear);";
```

```
proc recommend port=&lasrport;
add / item=movieid
user=userid
rating=rating;

addtable mylasr.ml100k / type=rating
vars=(movieid userid rating);

method svd / factors=10 maxiter=20
maxfeval=100 tech=lbfgs seed=1234
function=L2 lamda=0.1 label="svdlbfgs"
details;

method knn / similarity=pc positive k=20
label="knn1";

method ensemble / details label="em1"
maxiter=10 seed=4321;
run;
```

Thank You!

3 short years and…

- Hadoop is on the opening session
- 50% of the people attending an executive conference talk have an active Hadoop project
- SAS has built a family of applications that take full advantage of Hadoop as an analytics platform

Paul.Kent@SAS.com

@hornpolish

paulmkent

Washington, D.C.

March 23 – 26, 2014