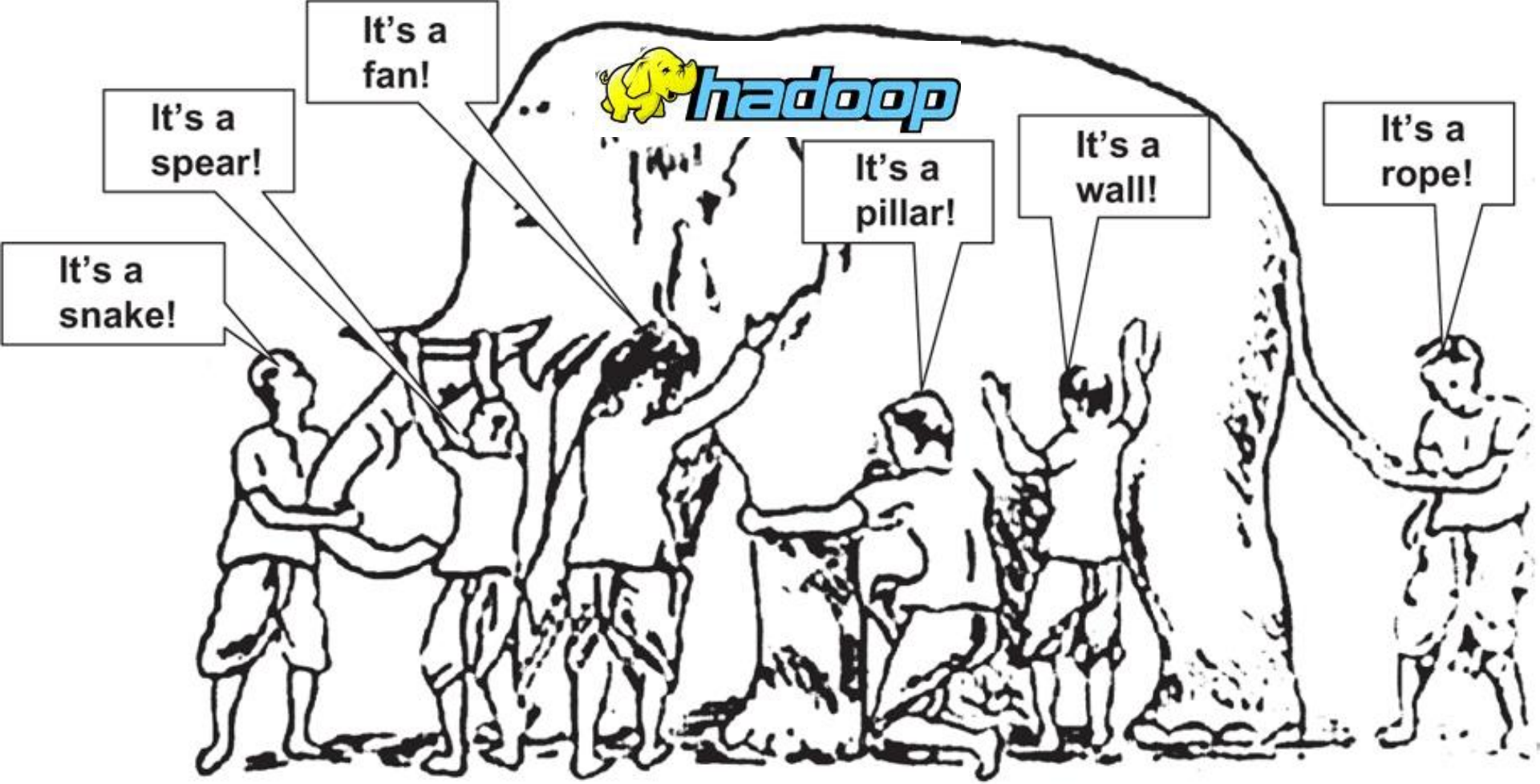
SAS® Forum
Nederland2014

Big Data Analytics met Hadoop

Jos van Dongen
Arno Klijnman

THE POWER TO KNOW.


It's a fan!
hadoop
It's a spear!
It's a pillar!
It's a wall!
It's a rope!
It's a snake!

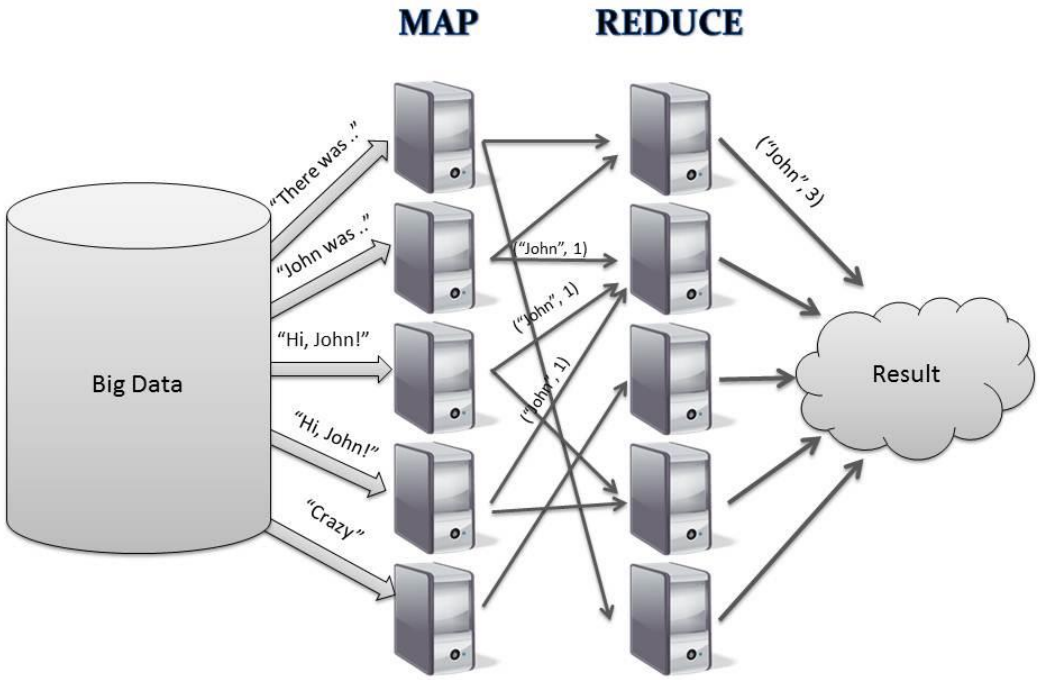
What is…

Distributed storage and processing of (big) data on large clusters of commodity hardware

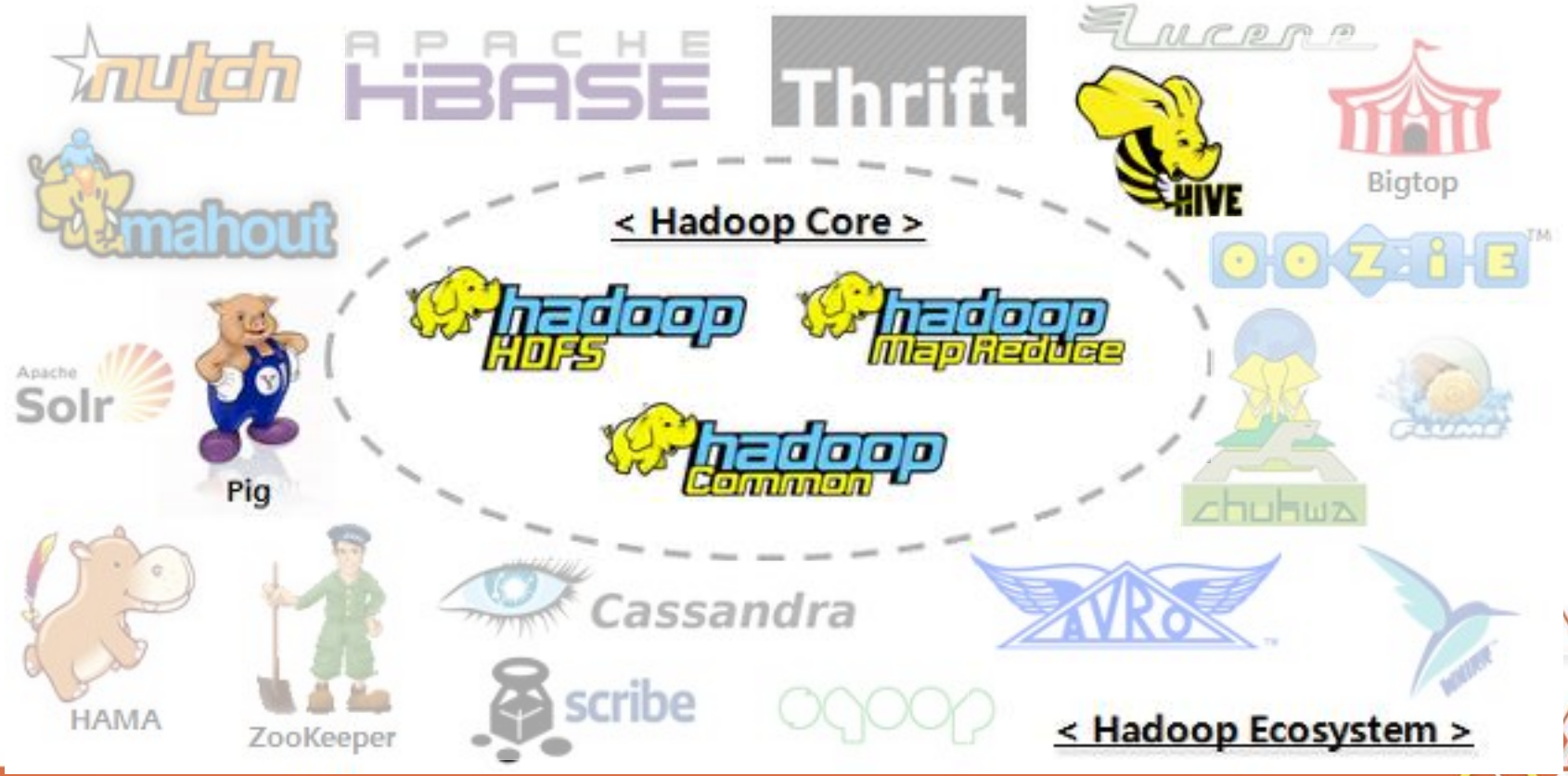
- HDFS
- Map/Reduce

HDFS - Distributed storage for big files

Map/Reduce- Distributed processing for big data

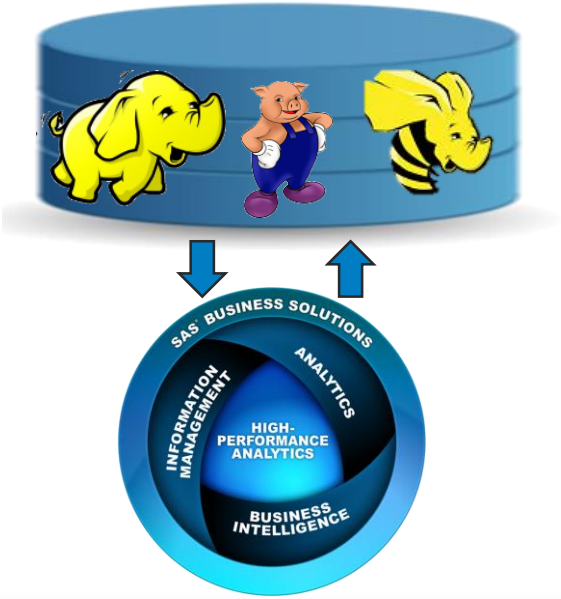


The Hadoop Jungle

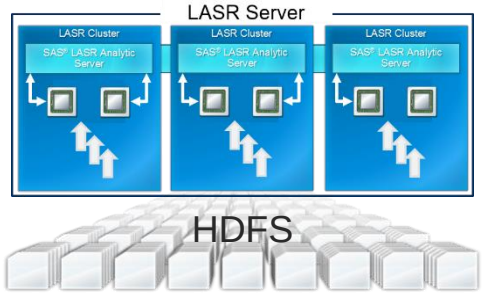


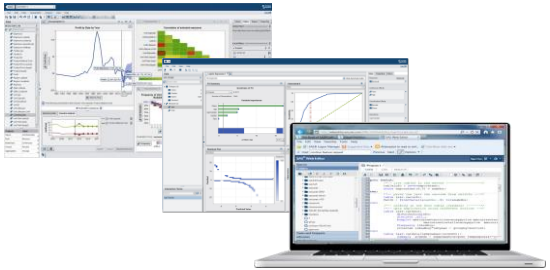
SAS & Hadoop Capabilities

WITH Hadoop

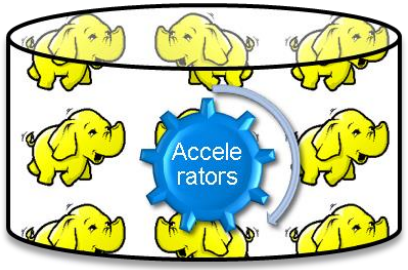


ON Hadoop



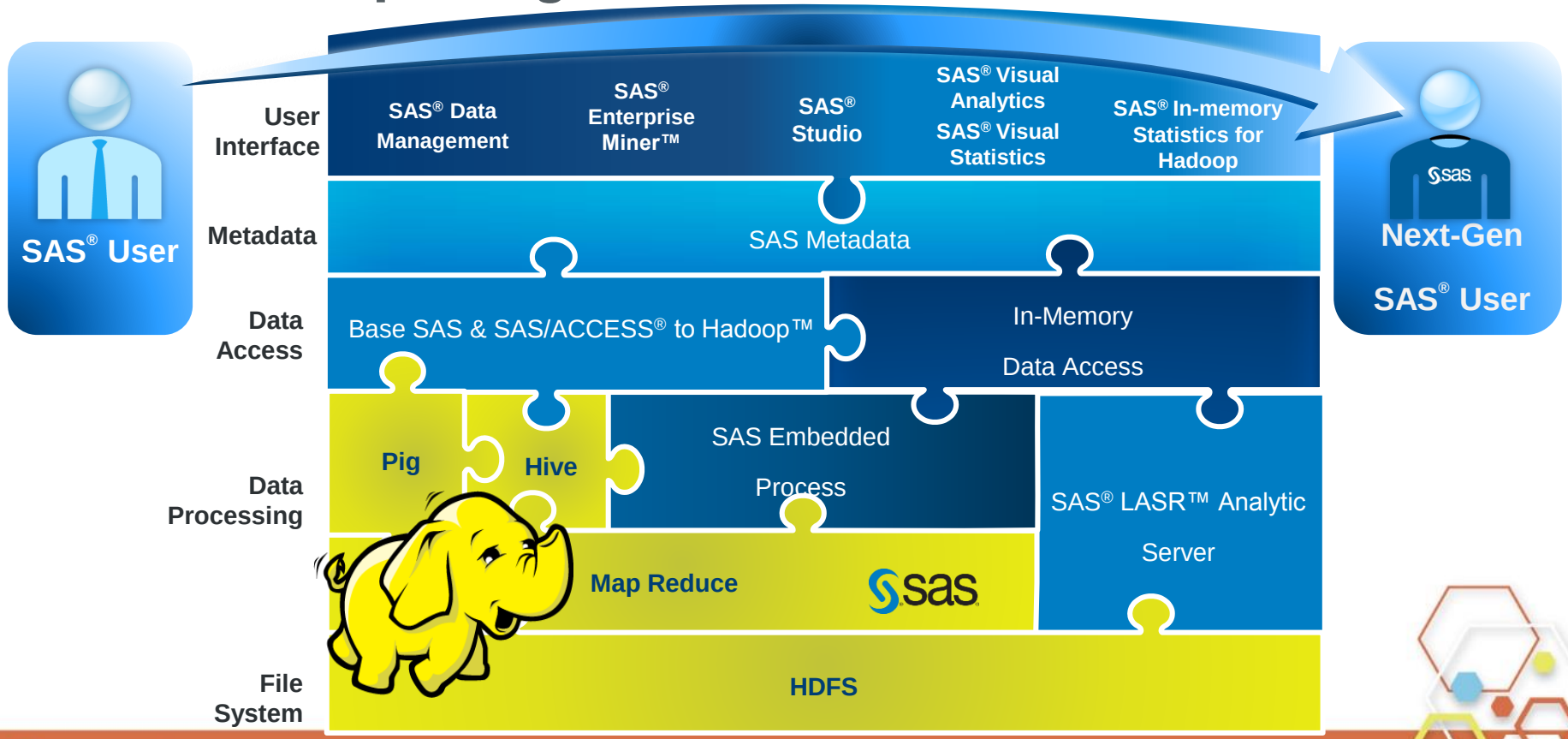


IN Hadoop



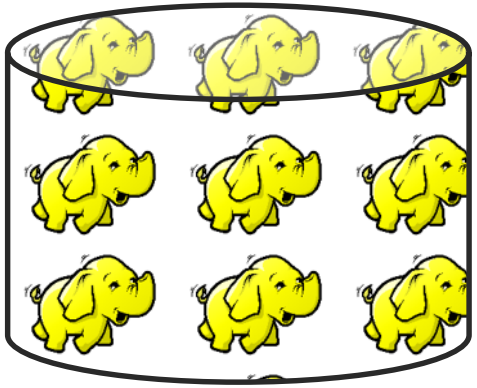
- **SAS Data Quality Accelerator**
- **SAS Scoring Accelerator**
- **SAS Code Accelerator**

SAS & Hadoop Integration

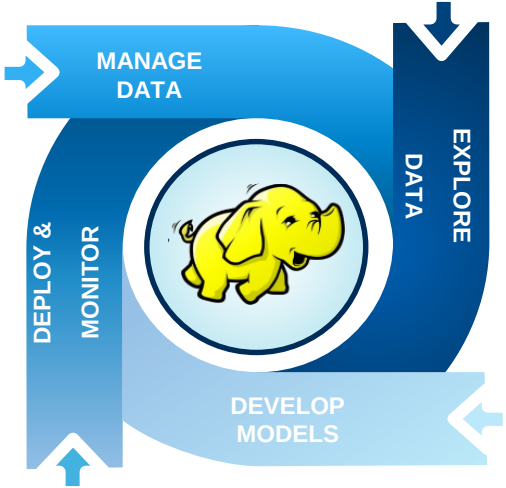


Two Paradigms

Hadoop as a Data Platform



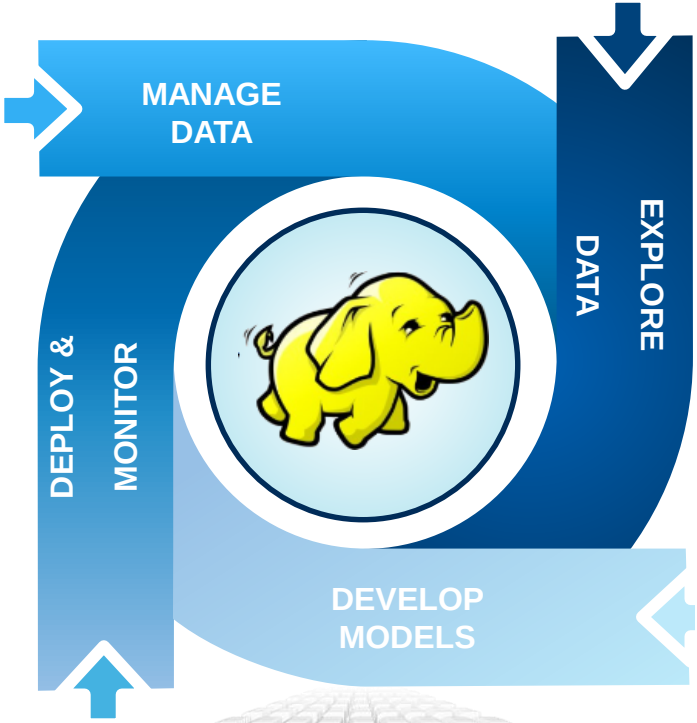
Hadoop as a core component of next generation analytical platform



Paradigm two

Hadoop as a core component of next generation analytical platform

- SAS/ACCESS
- SAS Data Management
- SAS Federation Server
- SAS Event Stream Processing
- SAS Data Loader for Hadoop
  - ✓ *SAS Data Quality Accelerator for Hadoop*
  - ✓ *SAS Code Accelerator for Hadoop*



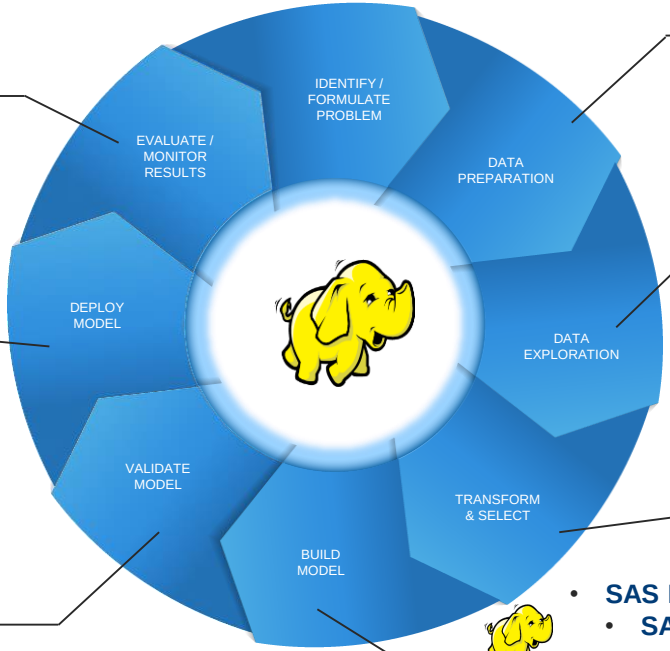
- SAS Data Loader for Hadoop
- SAS Visual Analytics
- SAS In-memory Statistics for Hadoop

- SAS Scoring Accelerator for Hadoop
- SAS Decision Manager
- SAS Visual Analytics

- SAS High Performance Analytics Products
- SAS Visual Statistics
- SAS In-memory Statistics for Hadoop

SAS runs the Entire Analytical Lifecycle in/on/with Hadoop

- **SAS Visual Analytics**



- **BASE SAS**
- **SAS / Access**
- **SAS Data Loader for Hadoop**
- **SAS DI Studio**

- **SAS Visual Analytics**
  SAS Visual Statistics
- **SAS High Performance Analytics Offerings**
- **SAS In-Memory Statistics for Hadoop**

- **SAS Scoring Accelerator for Hadoop**
- **SAS Code Accelerator for Hadoop**

Done using either the Data Preparation, Data Exploration or Build Model Tools

Done using the Build Model Tools and other checks

- **SAS High Performance Analytics Offerings**
- **SAS In-Memory Statistics for Hadoop**
- **SAS Visual Statistics**

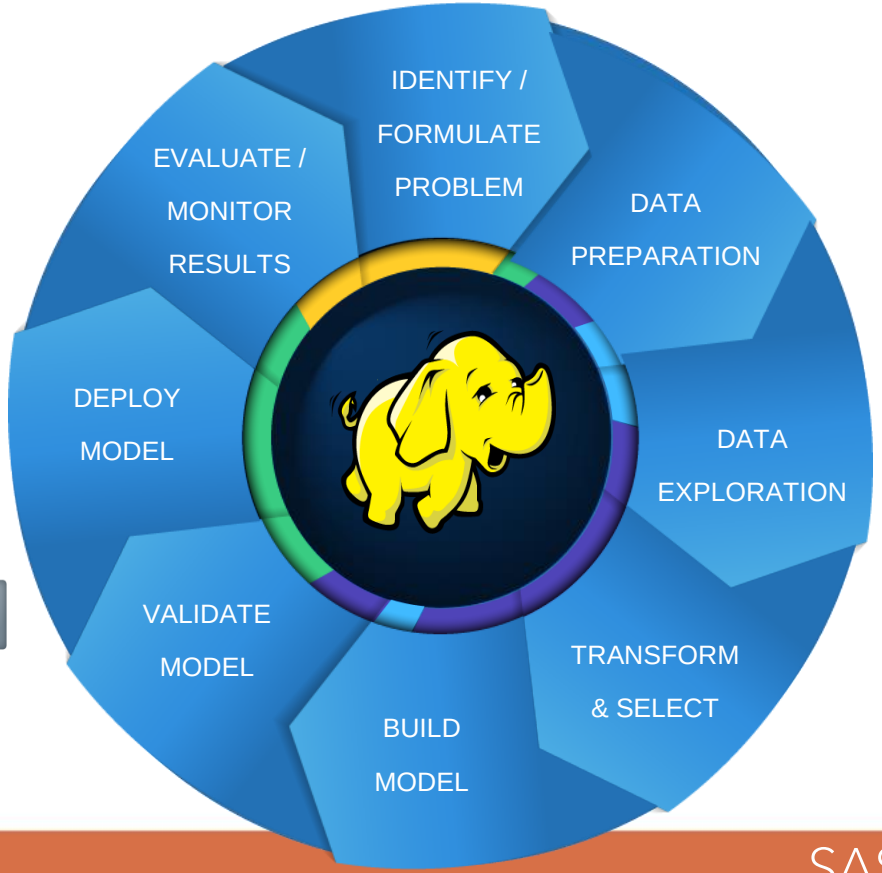
USER ROLES & THE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation



IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Data Preparation

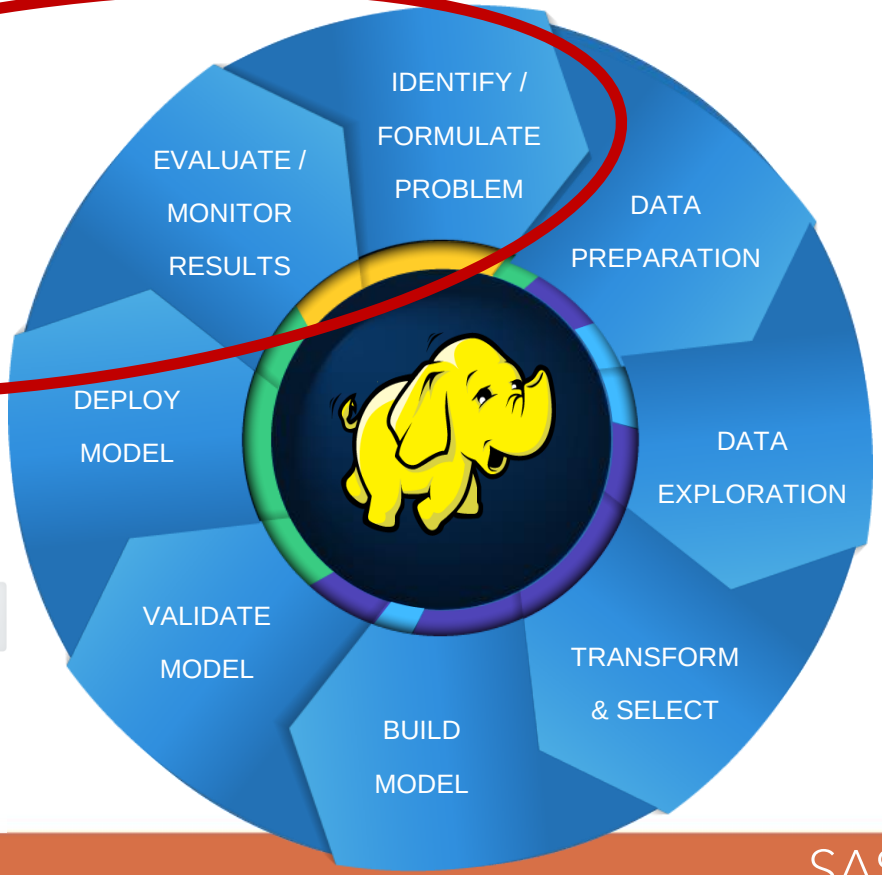
ANALYST DATA SCIENTIST

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

USER ROLES & THE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI



BUSINESS ANALYST

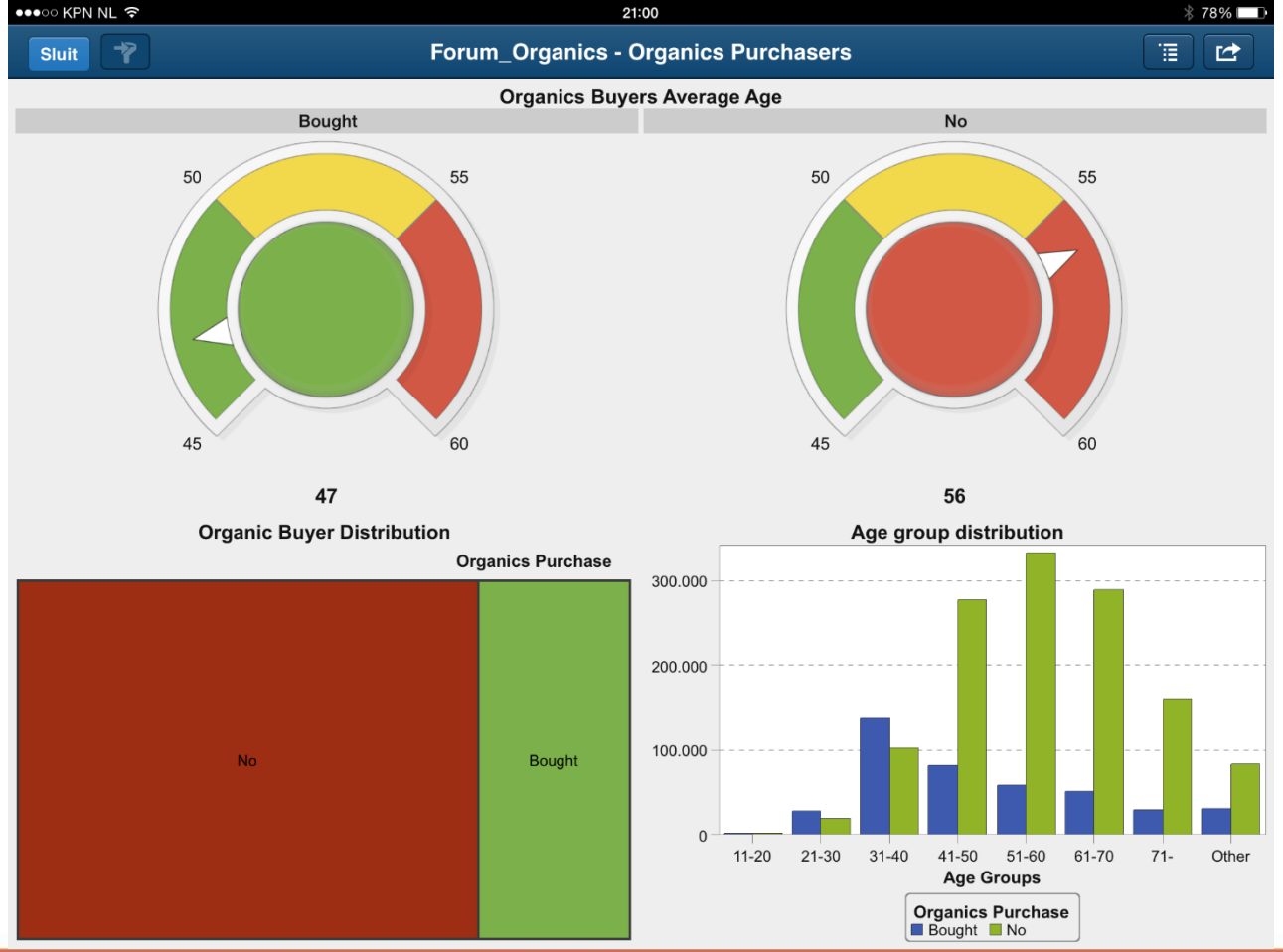
Data Exploration
Data Visualization
Report Creation

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Data Preparation

ANALYST DATA SCIENTIST

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling



SAS Forum Nederland2014

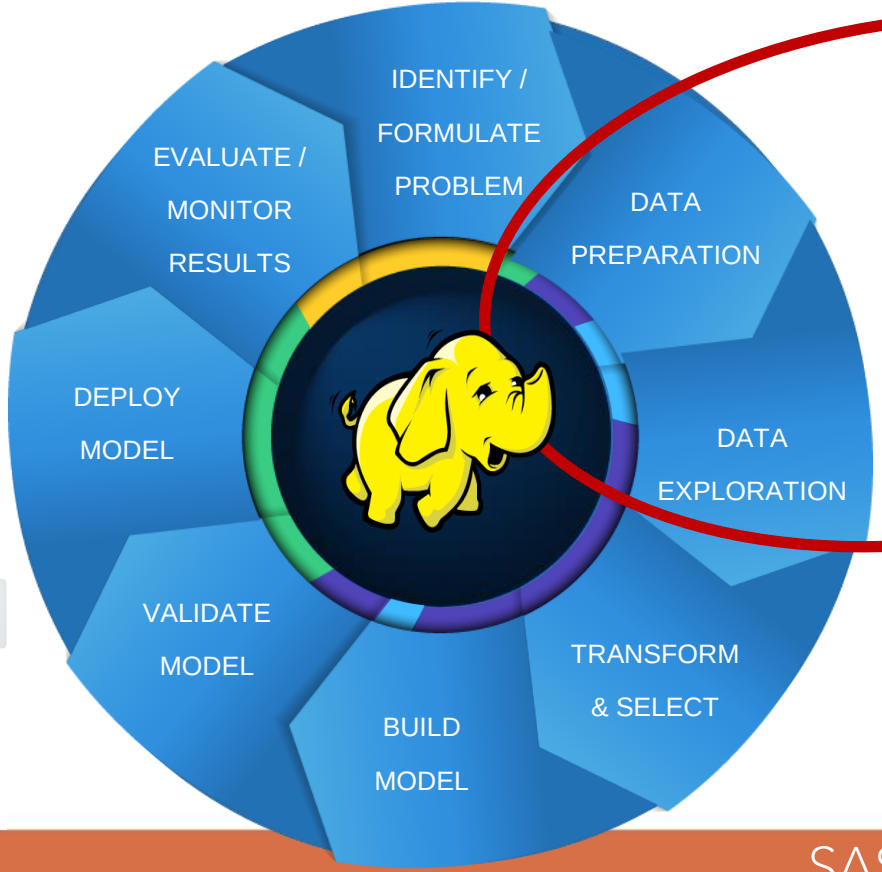
USER ROLES & THE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Data Preparation



BUSINESS ANALYST

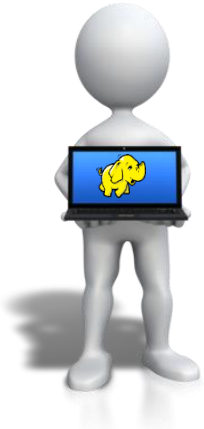
Data Exploration
Data Visualization
Report Creation

ANALYST DATA SCIENTIST

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

SAS® Data Loader for Hadoop

A new SAS Web-based Business user interface



Point & Click User Menus

Little or no Hadoop experience needed

Self-Service UI

HTML 5 Interface

What directive do you want to perform?

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

Run Status
Show the status of current and previous directive executions

Query a Table in Hadoop
Create a table by querying the data in an existing Hadoop table

Run a SAS Program
Run in-database data quality SAS programs

Transform Data in Hadoop
Transform data from a Hadoop table

Load Data to LASR
Copy data from a source and load it into LASR. Existing data in the target table will be replaced

Profile Data
Generate a profile report of the data in a table

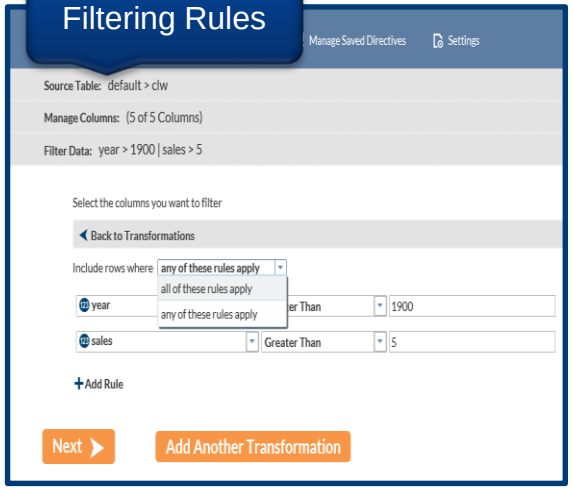
Saved Profile Reports
Explore previously generated profile reports

Enables Self-Service approach to managing data in Hadoop environment

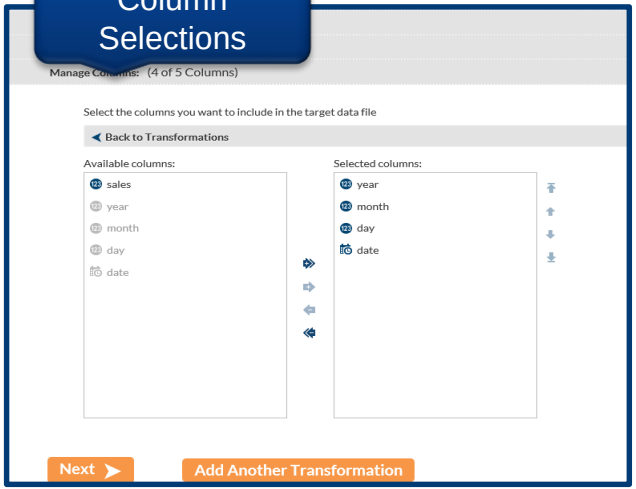
SAS® Data Loader for Hadoop

Transform Data in Hadoop

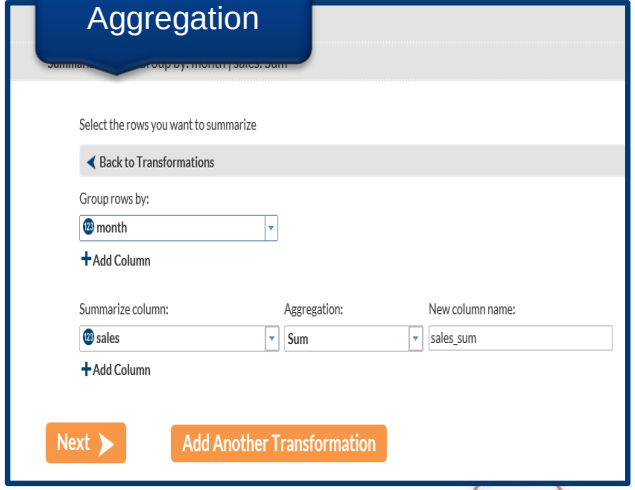
Filtering Rules



Column Selections



Aggregation

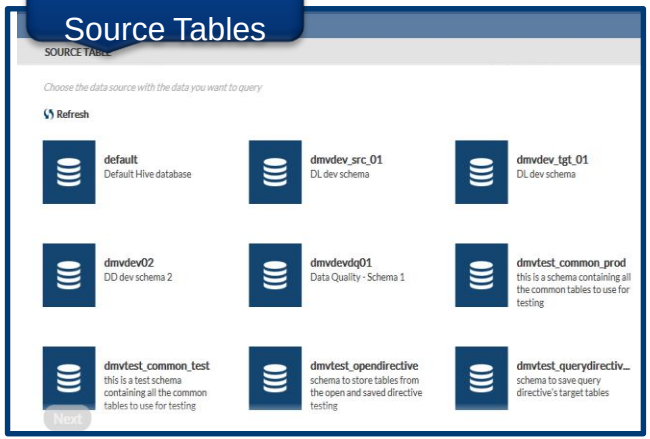


No coding, scripting or specialized skills required

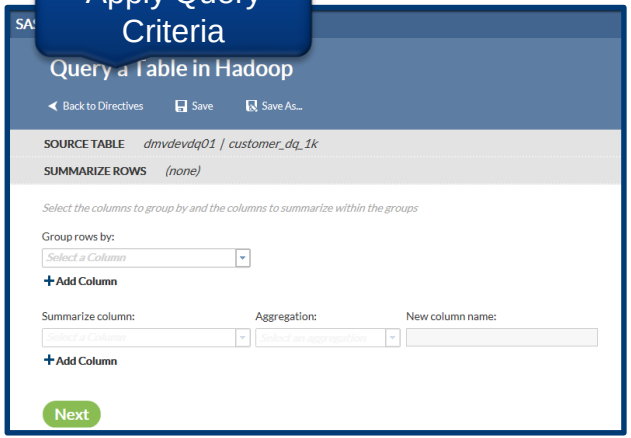
SAS® Data Loader for Hadoop

Query Hadoop data

Select Source Tables



Apply Query Criteria



See subset of data in Table Viewer

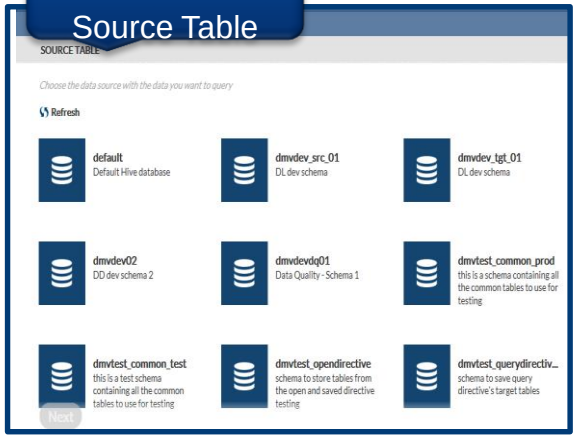
Simple Drag & Drop approach to Query Data inside Hadoop

SAS® Data Loader for Hadoop

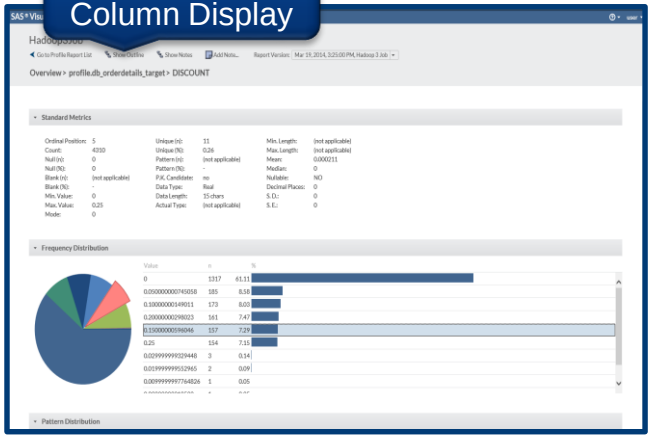
Profile Hadoop Data

Select

Source Table



View Reports in Column Display



View Reports in Table Display

Run standard metrics on data inside Hadoop and generate reports

SAS® Data Loader - Profile Reports

Organics1

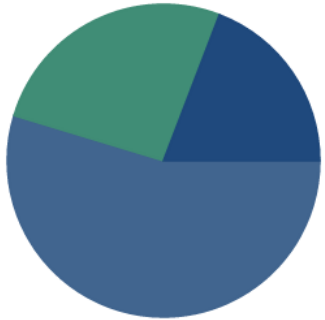
Go to Profile Report List | Show Outline | Show Notes | Add Note... | Report Version: Sep 26, 2014, 12:20:00 PM

default.organics > demgender

Standard Metrics

Ordinal Position:	4	Unique (n):	3	Min. Length:	1
Count:	1688948	Unique (%):	0	Max. Length:	1
Null (n):	0	Pattern (n):	1	Mean:	(not applicable)
Null (%):	0	Pattern (%):	0.0	Median:	(not applicable)
Blank (n):	0	P.K. Candidate:	no	Nullable:	(not specified)
Blank (%):	0	Data Type:	String	Decimal Places:	(not specified)
Min. Value:	F	Data Length:	1 chars	S. D.:	(not applicable)
Max. Value:	U	Actual Type:	string	S. E.:	(not applicable)
Mode:	F				

Frequency Distribution



Value	n	%
F	923324	54.67
M	441940	26.17
U	323684	19.16

SAS ® Data Loader

Query a Table in Hadoop

Back to Directives | Save | Save As...

SOURCE TABLE *default | organics*

SUMMARIZE ROWS *Group by: demgender, targetbuy | purchase_12mon: Sum*

Select the columns to group by and the columns to summarize within the groups

Group rows by:

demgender ✕

targetbuy ✕

+ Add Column

Summarize column:	Aggregation:	New column name:
purchase_12mon	Sum	total_purchase_12mon

+ Add Column

Next

FILTER ROWS *All rows*

COLUMNS *Columns specified in "Summarize Rows"*

SORT *(none)*

TARGET TABLE *default | totals_organic*

CODE *(generated code)*

RESULT *Successfully queried data*

View Data

SAS® Data Loader - Table Viewer

Schema Name: default Table Name: totals_organic Row Limit: 100

Columns

- demgender
- targetbuy
- total_purchase_12mon

Property	Value
Index	0
Label	demgender
Length	1
Name	demgender
Type	varchar

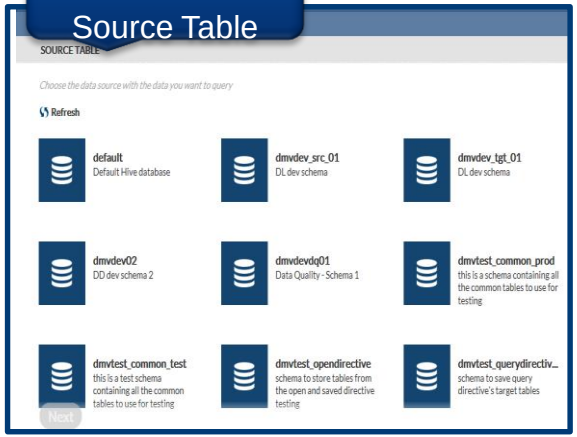
	demgender	targetbuy	total_purchase_12mon
1	F	Bought	1339040186.7499948
2	F	No	2529691969.1399984
3	M	Bought	307640939.4100002
4	M	No	1544105482.4699843
5	U	Bought	106367230.57000038
6	U	No	1249837525.04001

SAS® Data Loader for Hadoop

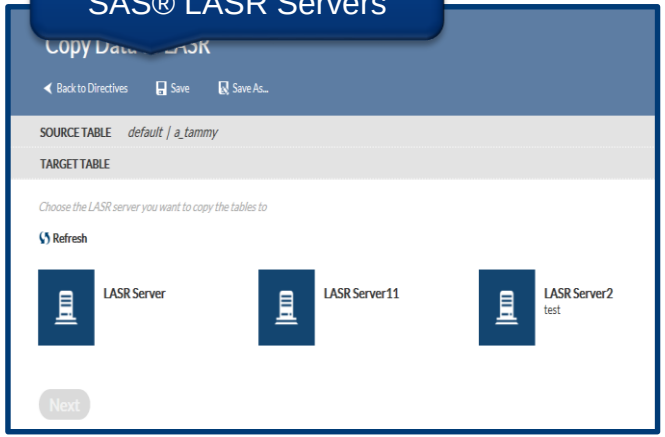
Copy Data to distributed sas® lasr server

Select

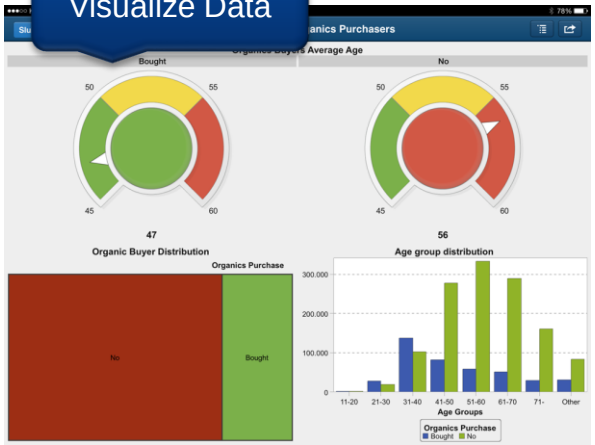
Source Table



Copy Data To distributed SAS® LASR Servers



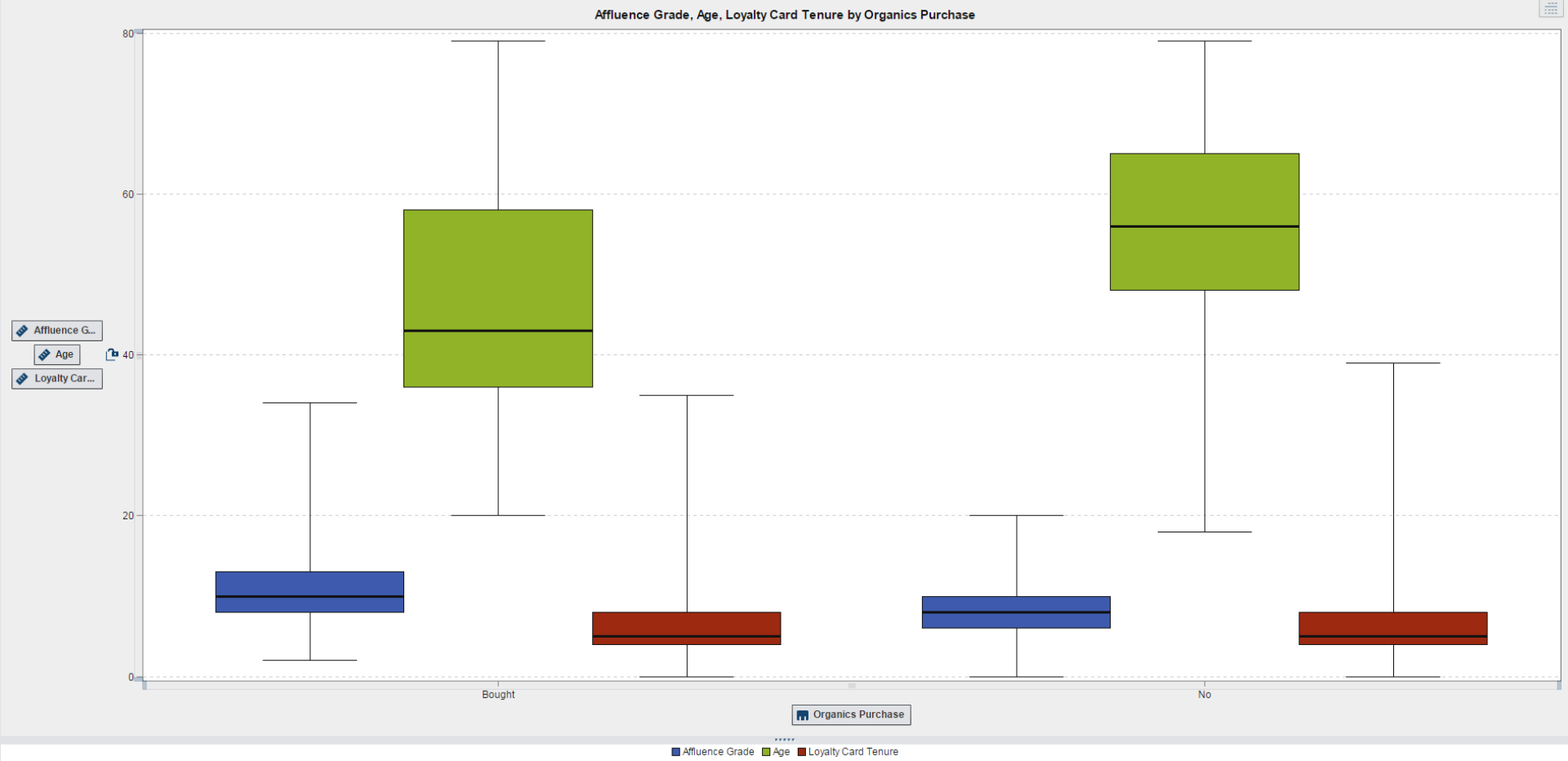
Visualize Data



SAS® Visual Analytics

Optional

Explore Hadoop data quickly and easily for faster insights



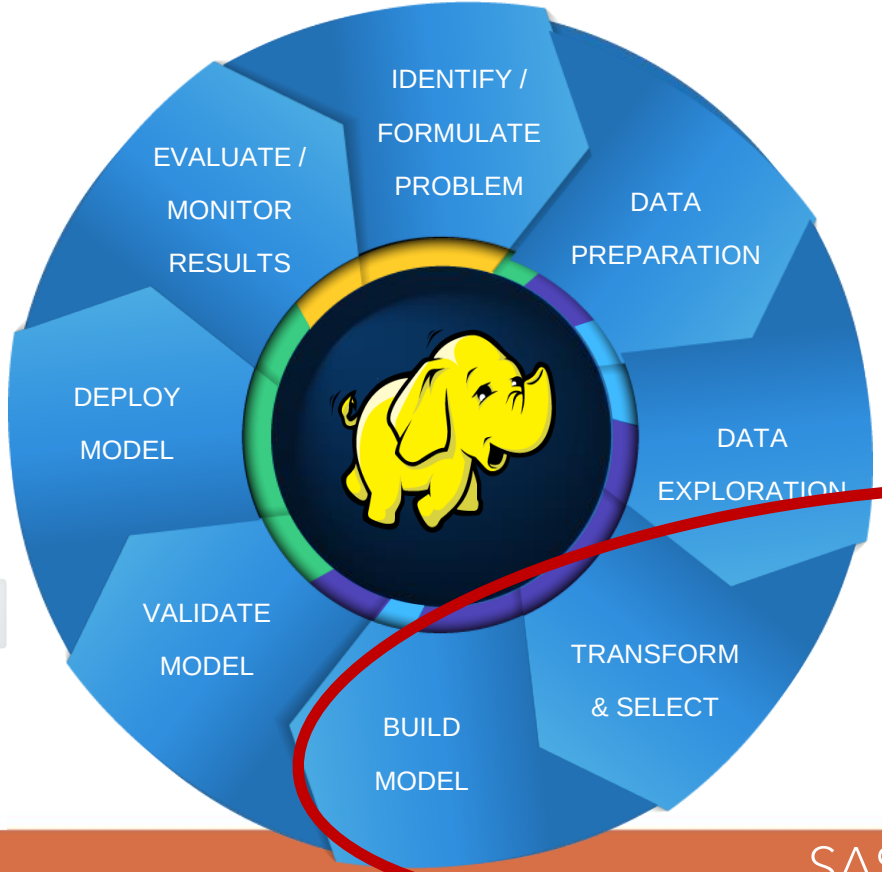
USER ROLES & THE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation



IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Data Preparation

ANALYST
DATA SCIENTIST

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Data

ORGANICS_NEW

Search data

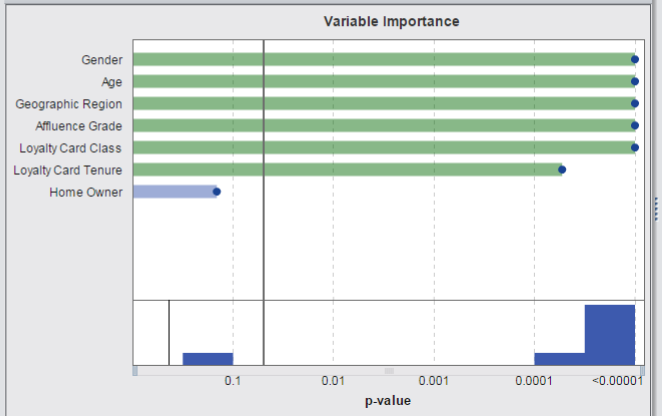
- Category (29)
  - Age Group 4
  - Bought Bakery 2
  - Bought Baking 2
  - Bought Beverages 2
  - Bought Canned 2
  - Bought Cleaners 2
  - Bought Dairy 2
  - Bought Frozen 2
  - Bought Fruits 2
  - Bought Meat 2
  - Bought Others 2
  - Bought PaperGoods 2
  - Bought Vegetables 2
  - Customer ID 5,000+
  - Demographic Flag 1 2
  - Demographic Flag 2 2
  - Demographic Flag 3 2
  - Demographic Flag 4 2
  - Demographic Flag 5 2
  - Demographic Flag 6 2
  - Demographic Flag 7 2
  - Demographic Flag 8 2
  - Gender 3
  - Geographic Region 6
  - Home Owner 2
  - Loyalty Card Class 4
  - Organics Purchase 2
  - Residential Neighborhood 56
  - Television Region 14
- Measure (7)
  - Affluence Grade
  - Age
  - Loyalty Card Tenure
  - Recent 12 Month Purchase
  - Recent 3 Month Purchase
  - Recent 6 Month Purchase
  - Recent 9 Month Purchase

Logistic 1

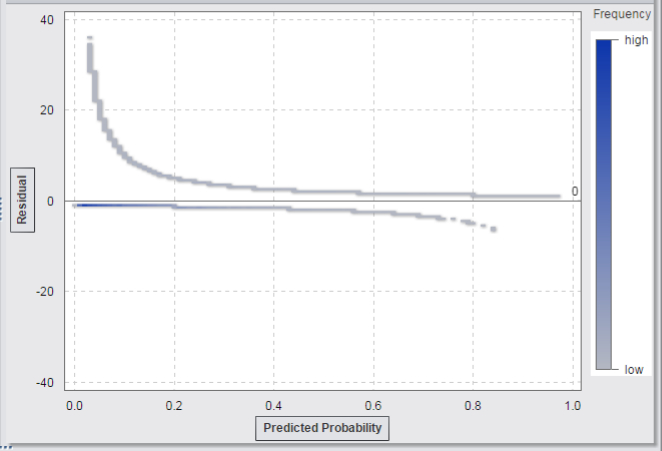
Organics Purchase (event=Bought) R-Square 0.1447 Observations: 1,176,988

Show Summary Table

Fit Summary

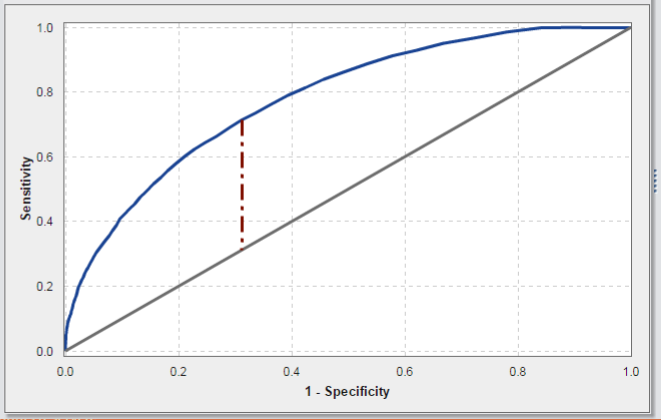


Residual Plot

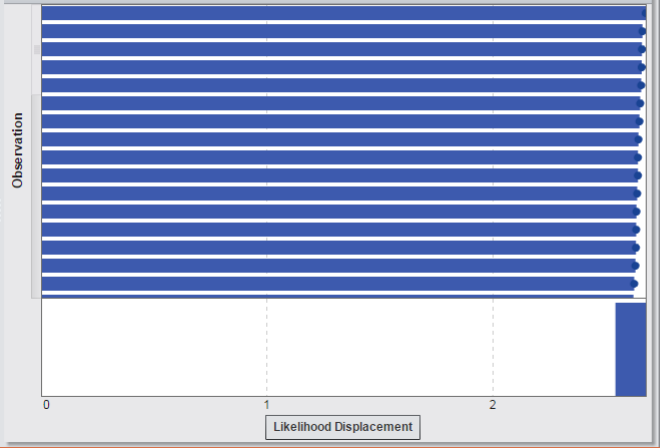


Assessment

ROC



Influence Plot



Roles | Properties | Filters

Logistic Regression

Response Advanced

Organics Purchase

Continuous Effects

- Affluence Grade
- Age
- Loyalty Card Tenure

Classification Effects

- Gender
- Geographic Region
- Home Owner
- Loyalty Card Class

Interactions Create

Term

Group By Advanced

Category

Frequency

Measure

Offset

Measure

Weight

Measure

Auto-update model

Update

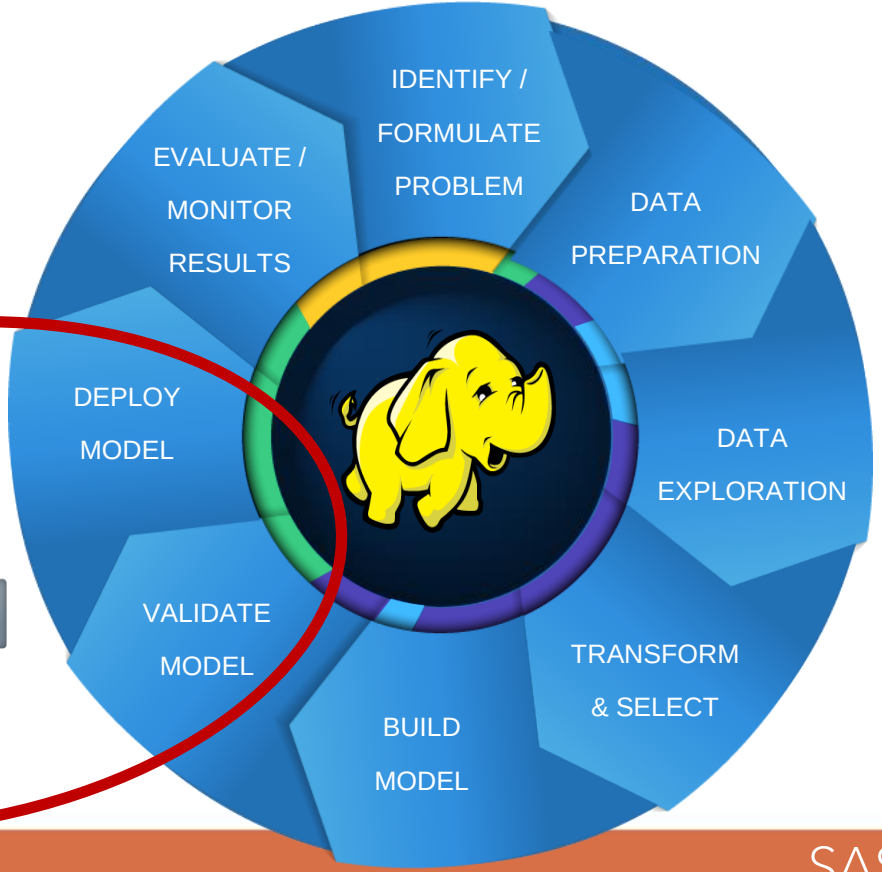
USER ROLES & THE ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Process and ROI

BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation



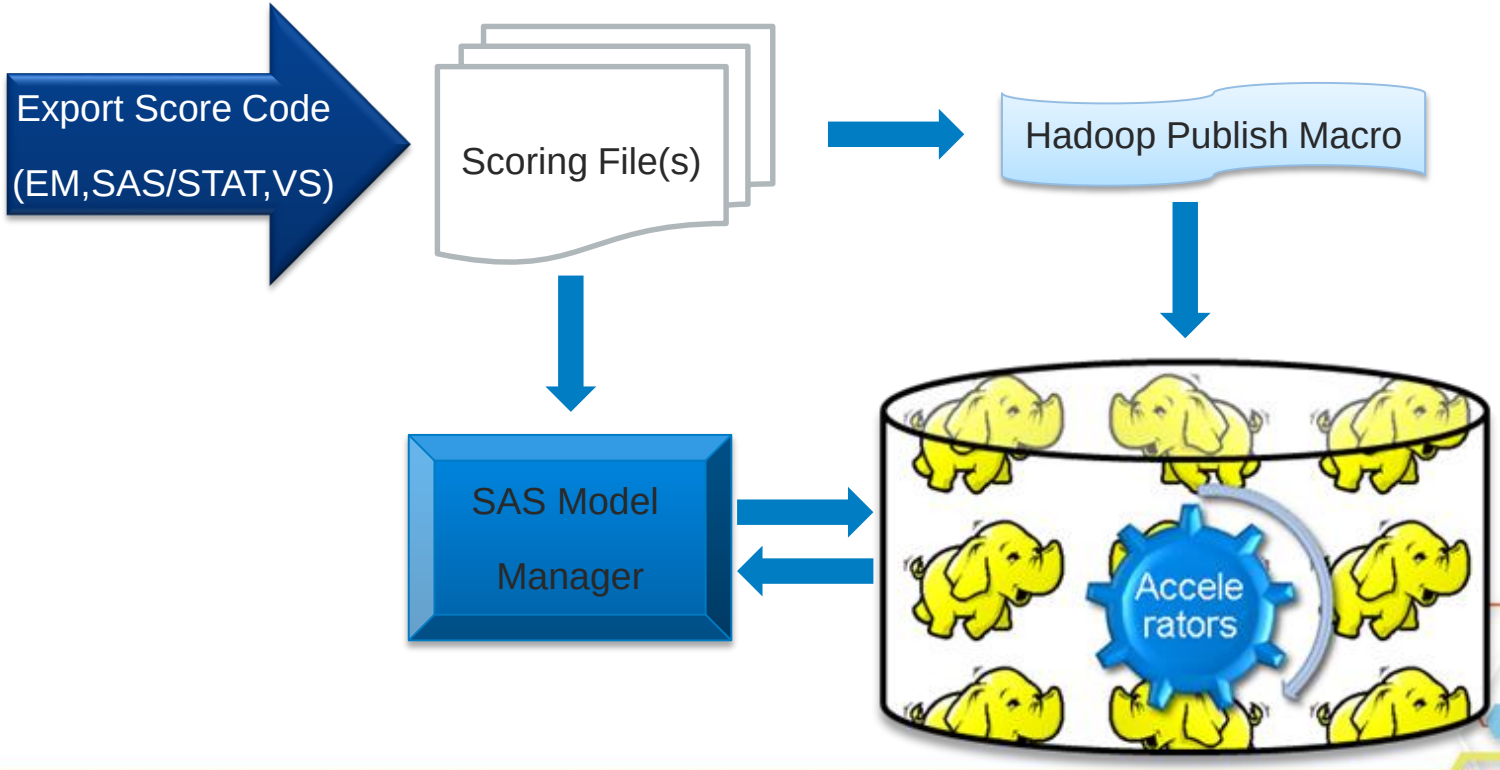
IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Data Preparation

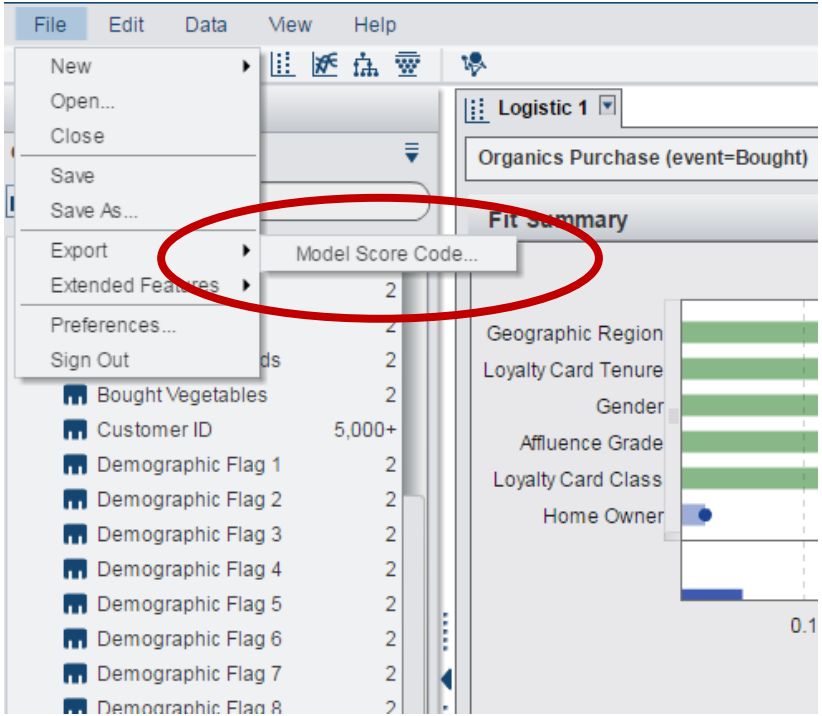
ANALYST DATA SCIENTIST

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

SAS Scoring Accelerator for Hadoop



SAS Scoring Accelerator for Hadoop



SAS - [LM_Organics.sas]

File Edit View Tools Run Solutions Window Help

Explorer

Contents of 'SAS Environment'

Libraries File Shortcut

Favorite Folders Computer is SNLJDO

```
/**** User [illegible] ****/

/**** Model Name: LM_Organics ****/
/*-----------------------------------------
SAS Code Generated by LASR Analytic Server
  Date              : 29Sep2014:08:53:17
  Model Type        : Logistic Regression
  Class variable    : DemGender
  Class variable    : DemReg
  Class variable    : DemHomeowner
  Class variable    : PromClass
  Response variable : TargetBuy
  Distribution      : Binary
  Link Function     : Logit
  -----------------------------------------*/

/*-----------------------------------------*/
/*Defining temporary arrays and variables   */
array _xrow_0_{19} _temporary_;
array _beta_0_{19} _temporary_ (   -4.607037[illegible]587436
        0.27348284016585
        -0.[illegible]
        -0.0023694971714
        1.97324667117803
        0.99682178251566
                       0
       -0.07057240838864
       -0.20582048166369
       -0.13540275941414
       -0.13956928605968
        0.06945462962515
                       0
        0.00837088260452
                       0
       -0.02034794541427
       -0.31929029038522
       -0.09617084185945
                       0);

drop _badval_ _linp_ _temp_ _i_;
_badval_ = 0;
_linp_   = 0;
_temp_   = 0;
_i_      = 0;
```

Output - (Untitled) | Log - (Untitled) | LM_Organics.sas

Results | Explorer

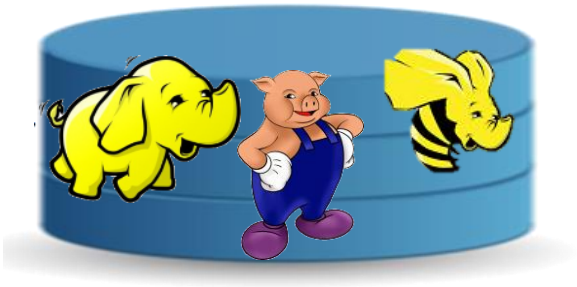
Autosave complete | C:\Users\snljdo | Ln 45, Col 1

Demo flow

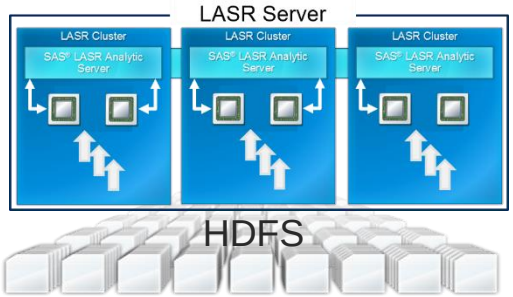
SAS DI	Data Loader	VA Explorer	VS	IMSTAT	Scoring Accelerator
• Access to Hadoop • Transform • Write back to Hadoop • Write to LASR	• Show table • Profile • Build Query • Write result to LASR	• Discover relations • Understand the data	• Discover a model • Determine significance • Cluster variables	• Recommendation • Datastep to enrich original dataset with recommendation results • Write to LASR	• Deploy model • Run model • Back to Data Loader
SAS Data Management		SAS Interactive Analytics On Hadoop			SAS Analytics

SAS & Hadoop: 3 Things to Remember

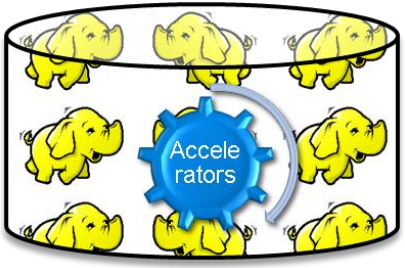
WITH Hadoop



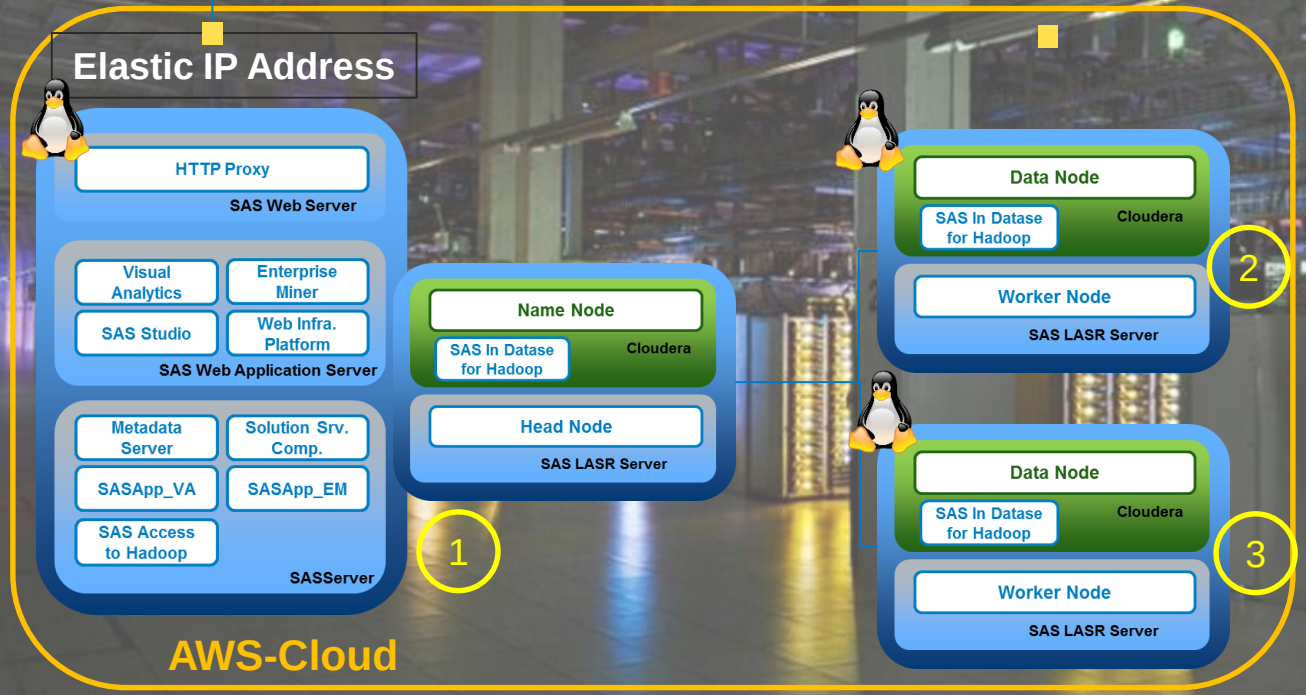
ON Hadoop



IN Hadoop



Demo Environment Infrastructure



Setup:

- CentOS operating system
- Local users on all Amazon servers
- Internal network for all Amazon Servers
- Open firewall for all ports between workstation & server
- No integration Mail server
- No SSL

SAS® Solutions for Hadoop

Hadoop brings big data. We bring everything else.

View video

Feedback

THE POWER TO KNOW.

http://www.sas.com/nl_nl/events/14/Masterclasses/Datasience.html

Alle opleidingen: Gesorteerd op rol | Opleidingskalender

More value from Big Data with Hadoop and SAS

ZIE STARTDATA EN REGISTREER

9 oktober 2014

Huizen

More value from Big Data with Hadoop and SAS

Greater analyst productivity and power to solve complex problems by uncovering undetected patterns and trends faster than ever with SAS and Hadoop.

SAS software for Hadoop enables multiple users to concurrently manage and prepare data stored in Hadoop, explore and visualize this data, develop accurate statistical and machine learning models quickly, as well as access, deploy and execute these models in their Hadoop ecosystem.

SAS empowers organizations to start exploring the value of big data in Hadoop rather than just collecting and storing the data. Statisticians and Data Scientists can address all steps of the analytical lifecycle, going from raw Hadoop data to integrating analytical models into business processes. By using state-of-the-art statistical algorithms and machine-learning techniques, multiple users can concurrently explore and use multiple analytic approaches to build models and quickly run multiple iterations to determine the best models. This vastly improves end-user productivity of scare data scientist resources, increases model building efficiency, decreases time from model inception to deployment and generates faster time to insights for making better decisions. As a result, organizations can now quickly and efficiently delve deep into Hadoop to make decisions based on fact based analytical insights into all of the data for competitive advantage.

This session will focus on the data management and data preparation steps needed to fuel the analytical capabilities of your SAS on Hadoop solution, whether it is SAS Visual Analytics, SAS Visual Statistics or SAS In Memory Statistics for Hadoop. You will learn which solutions are available to manage data in Hadoop, prepare data for In-Memory analysis and the various options there are for managing data quality, model scoring and executing SAS code, all running natively in a Hadoop cluster.

SAS® Forum Nederland 2014

THE POWER TO KNOW.