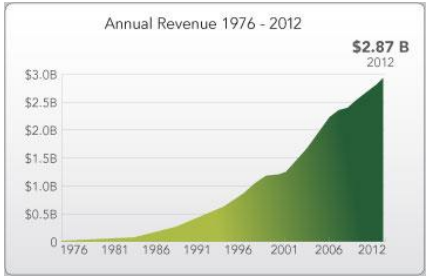
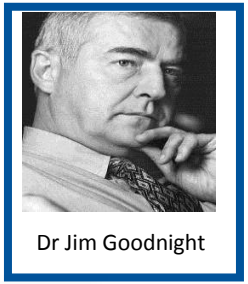
OVERVIEW SAS APPROACH ON BIG DATA ANALYTICS

MAURIZIO SALUSTI SAS

SAS OVERVIEW

37	Years since the founding
13.500	Employees
1	Business Analytics Provider
60.000	Installations worldwide
25%	revenue on R&D
96	Customers on Fortune 100
98%	Global Companies consider first 2.000
135	Countries





Dr Jim Goodnight

GREAT PLACE TO WORK

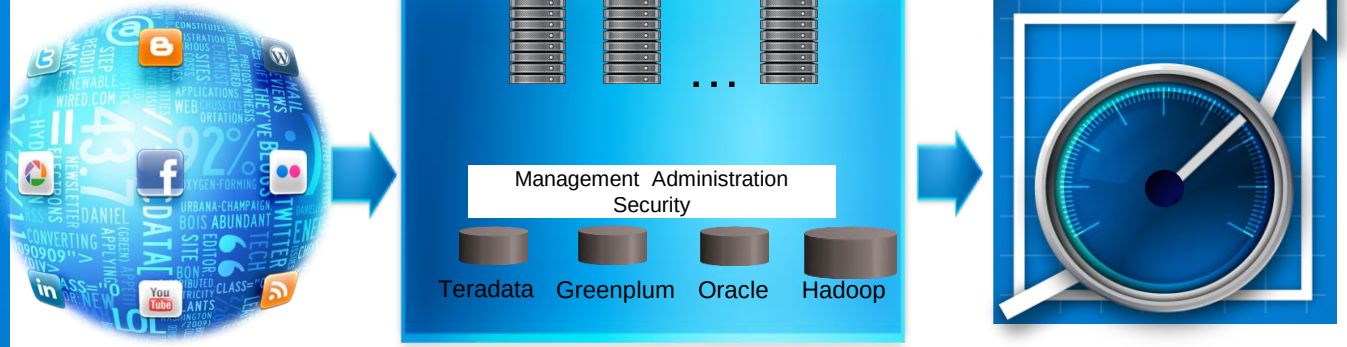
SAS® HIGH-PERFORMANCE ANALYTICS PRODUCTS

BIG DATA IS INFLUENCING INFORMATION ARCHITECTURE FOR ANALYTICAL MODELING

All of your data

Model extensively, iteratively, frequently

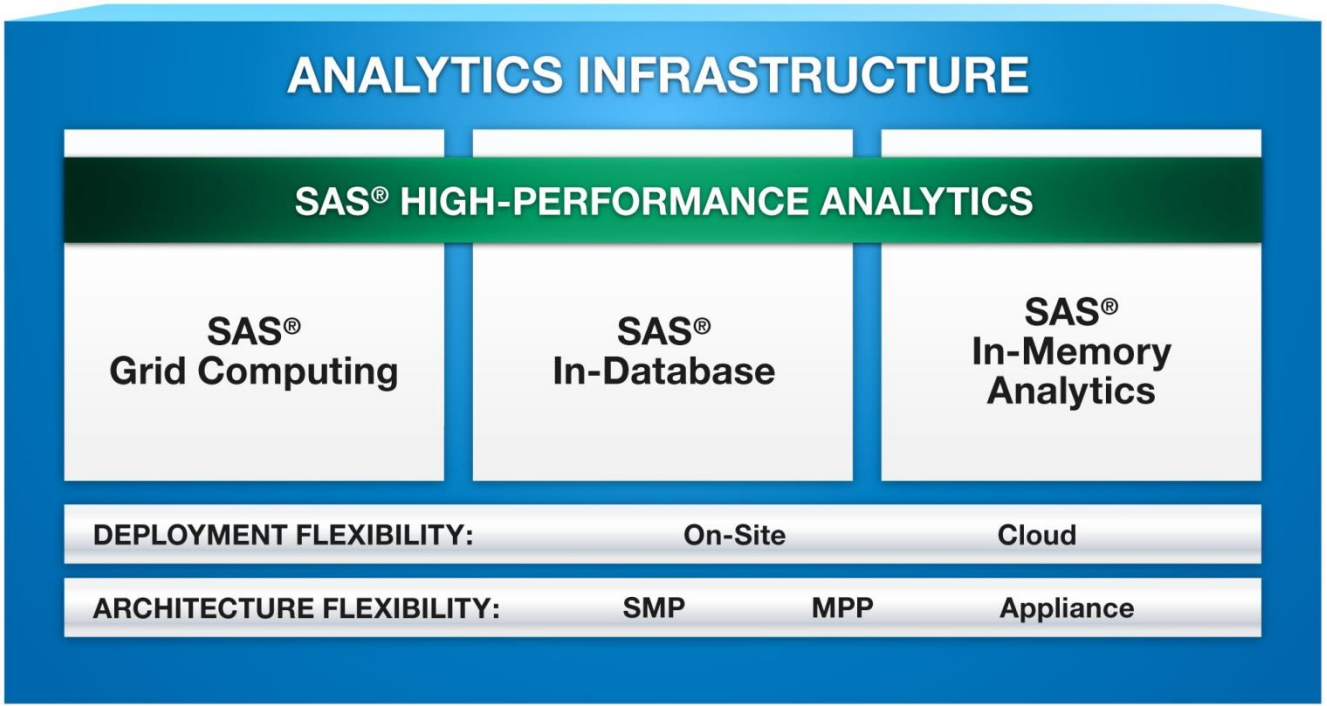
Better decisions all the time



sas THE POWER TO KNOW.

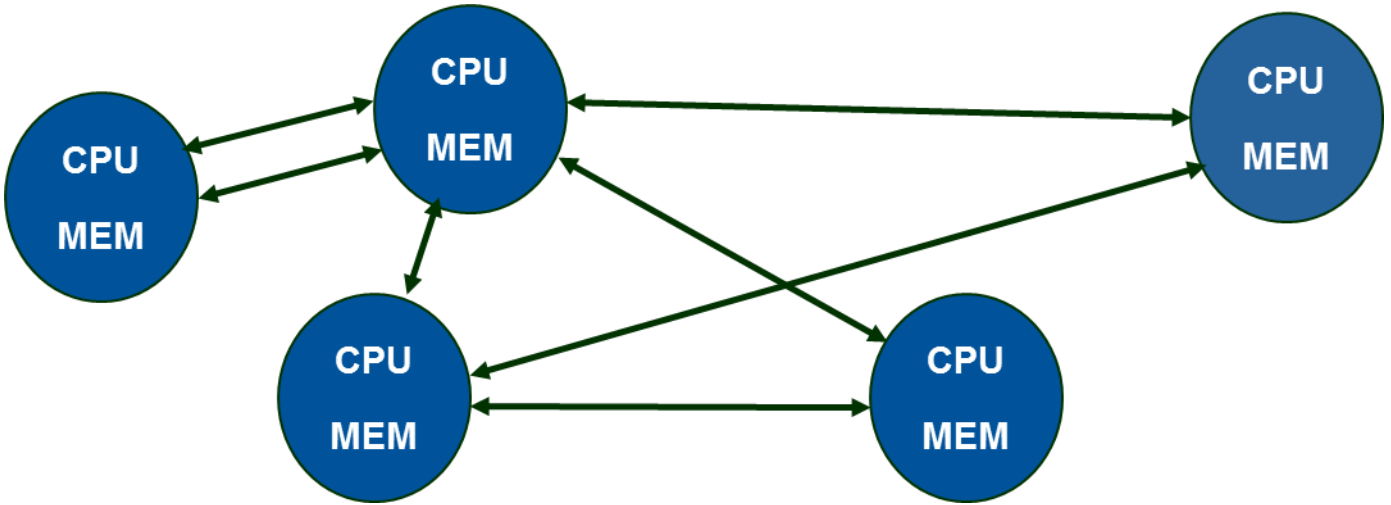
SAS® HIGH-PERFORMANCE ANALYTICS

KEY COMPONENTS



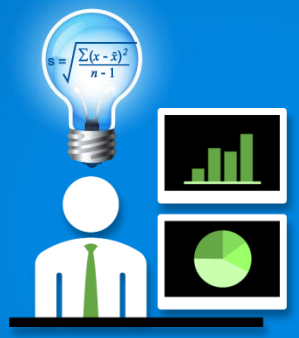
MESSAGE PASSING MODEL (SHARED NOTHING)

- DISTRIBUTED SET OF PROCESSORS EACH WITH IT'S OWN MEMORY (I.E. PRIVATE DATA)
- INTERCONNECTION NETWORK
- ALL **COMMUNICATION** AND **SYNCHRONIZATION** IS PERFORMED THROUGH THE EXCHANGE OF MESSAGES DIRECTLY **IN MEMORY** BETWEEN ALL **GRID NODES**

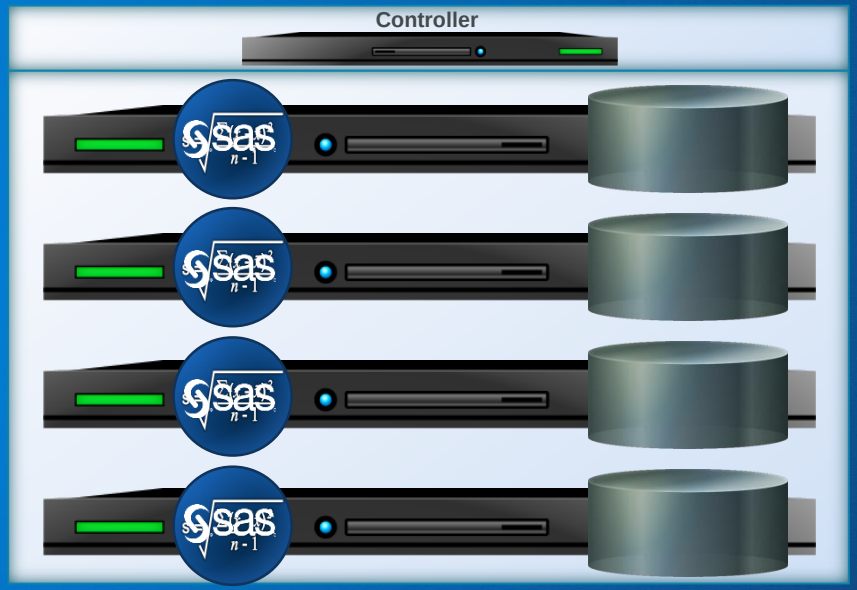


SUPPORTED HIGH-PERFORMANCE ARCHITECTURE

SAS PROCESSING IN-TANDEM WITH DATA

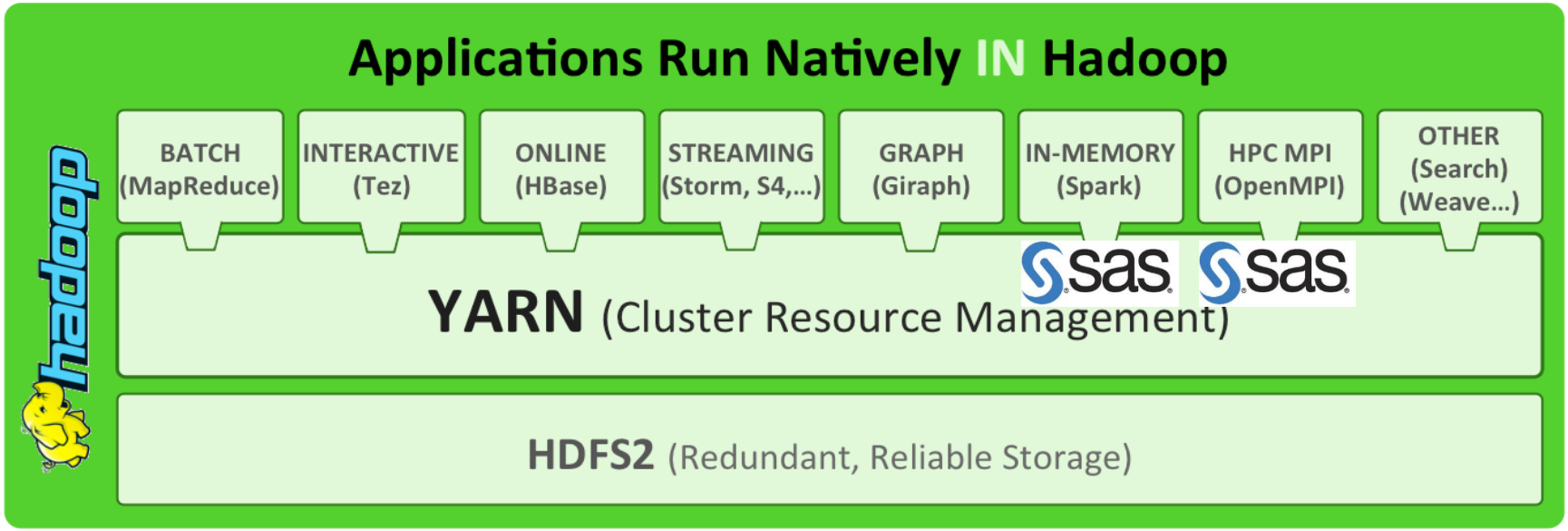


SAS® ANALYTICS Client

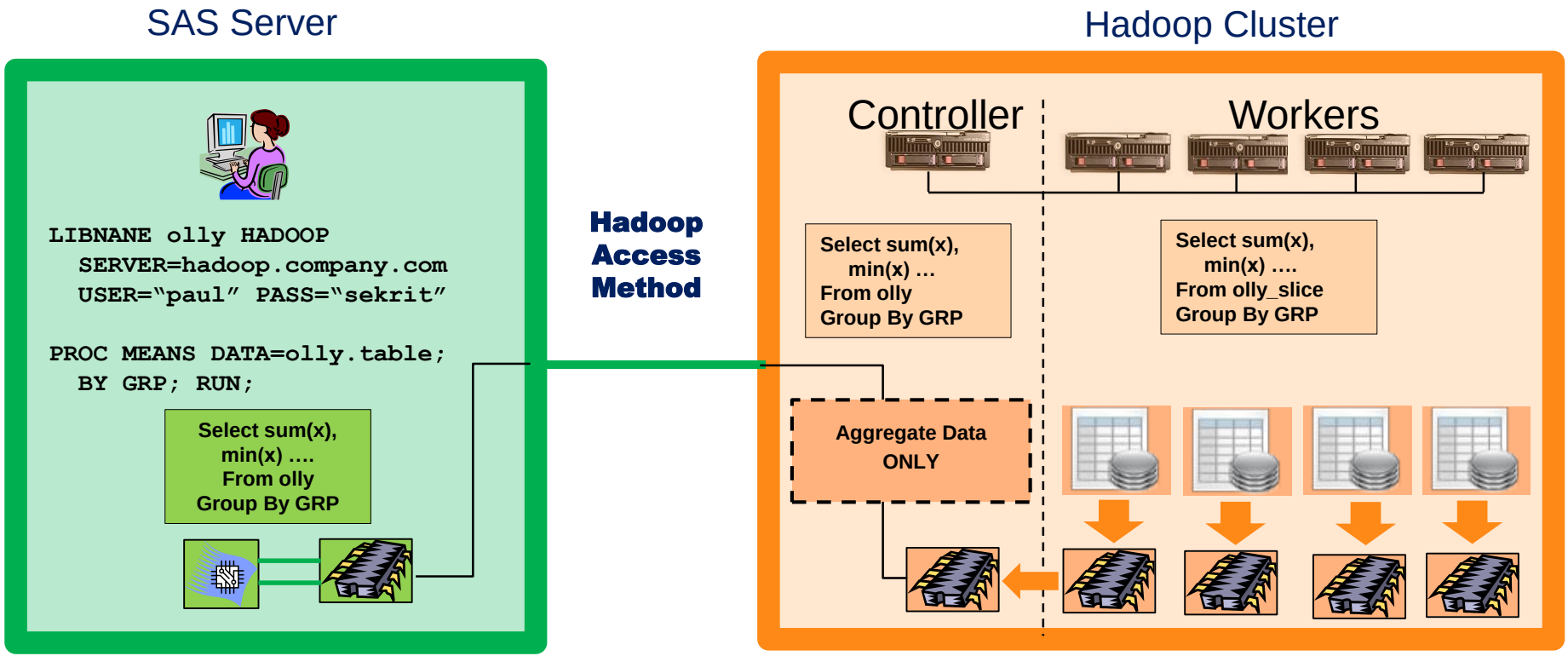


Apache Hadoop on Commodity Hardware

DIRECTION WITH HADOOP



USING SQL TO ACCESS DATA



USING MPI AND IN MEMORY CALCULATION

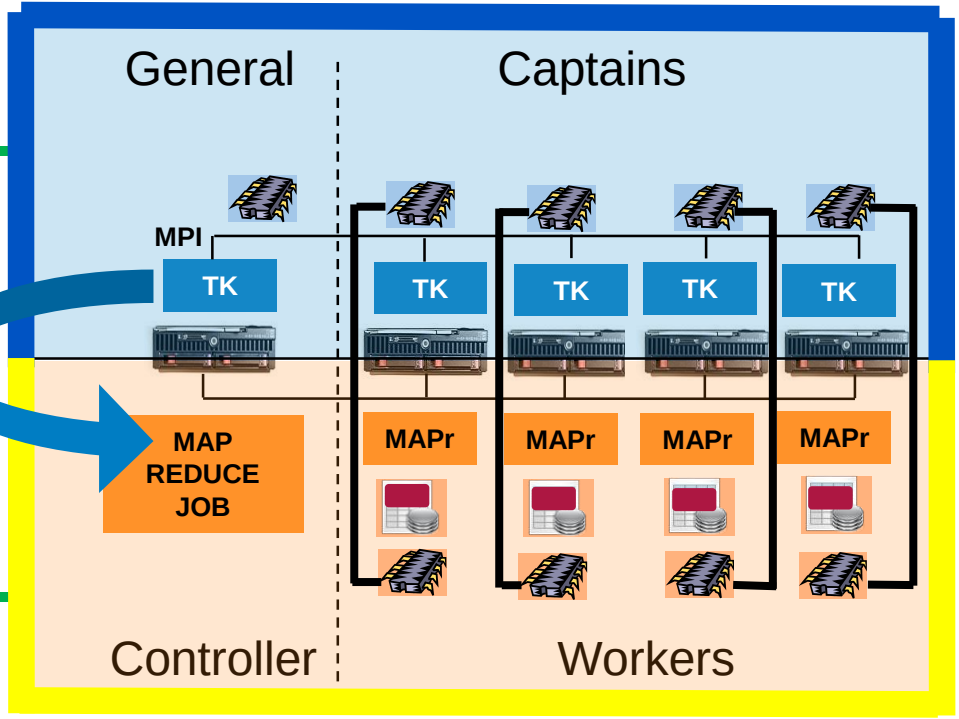
SAS Server

```
libname joe sashdat "/hdfs/..";

proc hpreg data=joe.class;

 class sex;
 model age = sex height
             weight;
run;
```

TKGrid

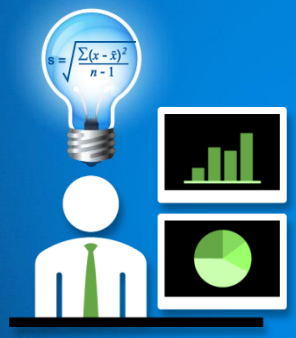
Hadoop Access Method

Appliance



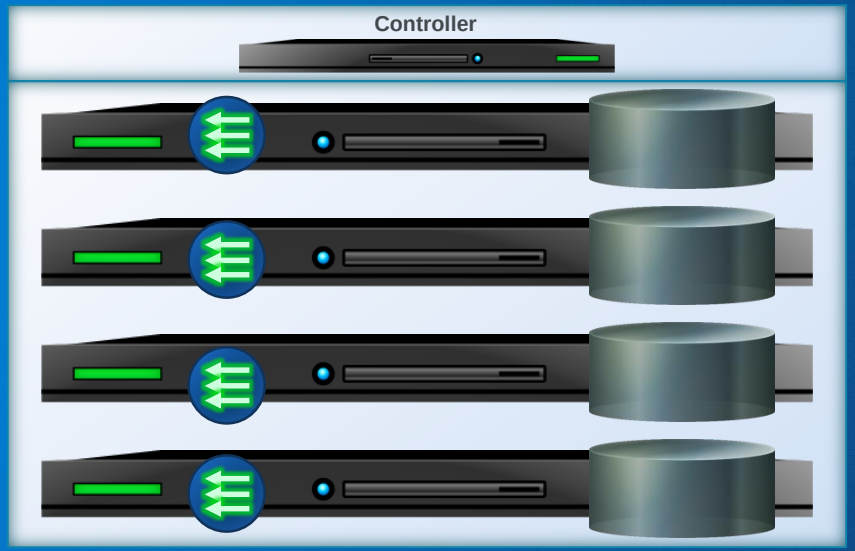
SUPPORTED HIGH-PERFORMANCE ARCHITECTURE

SAS PROCESSING DIRECTLY ATTACHED TO DATA



SAS® ANALYTICS Client

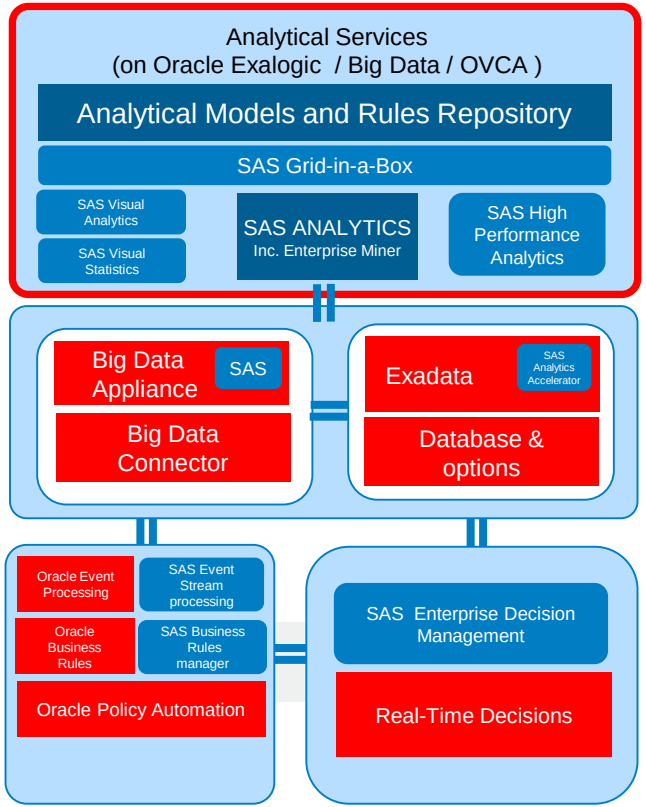
Compute Appliance



Existing Teradata or Oracle or Pivotal (Greenplum) Database or Hadoop

ANALYTICAL WORKLOAD

RAPID TIME TO VALUE IN STANDARD DEPLOYMENT



ORACLE SAS

- Using Different Data and Computing Appliances with Asymmetric HPA
-

SAS Server

```
libname a oracle
 server="dataAppliance";

proc hpcorr data=a.flights;
 performance
  mode=asym
  host="computingAppliance";
run;
```

TKGrid

Access Engine

Computing Appliance (Exalogic/BDA/OVCA)

General

Captains

TK

TK

TK

Data Appliance (Exadata)

Controller

Workers

SAS® HIGH-PERFORMANCE ANALYTICS PRODUCTS

BIG DATA IS INFLUENCING INFORMATION ARCHITECTURE FOR ANALYTICAL MODELING

- Data (“All data”, number of variables, new events, unstructured, …)
- Visualize (fast, interactive, analytical, evaluate, …)
- Models (no. of iterations, complex models, retraining, ensembles, …)
- Deploy (operationalize, real-time, in-database, …)

SAS® HIGH-PERFORMANCE ANALYTICS PRODUCTS

DESIGN PRINCIPLES

High-Performance Procedures

- Run parallel in single-machine mode (SMP) using concurrently scheduled threads.
- Use same syntax to run in a distributed mode (MPP) using multiple concurrently scheduled threads on each machine in a cluster.

Realize models in many complex situation

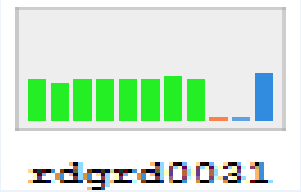
Open new way of analysis

- Perform variable selection and identification that generalize well for big data (billions of rows, thousands of variables).
- Focus on procedures for predictive and prescriptive modeling vs. computing inferential statistics on small data.
- Make faster very complex algorithm on large dataset

```
proc logistic data=TD.mydata;
   class A B C;
   model y(event='1') = A B B*C;
run;
```

```
proc hplogistic data=TD.mydata;
   class A B C;
   model y(event='1') = A B B*C;
run;
```

Single / Multi-threaded



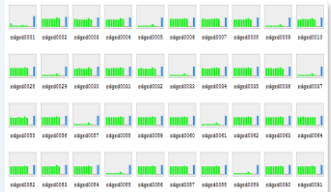
Not aware of distributed computing environment

Computes locally / where called

Fetches Data as required

Memory still a constraint

Massively Parallel (MPP)



Uses distributed computing environment

Computes in massively distributed mode

Work is co-located with data

In-Memory Analytics

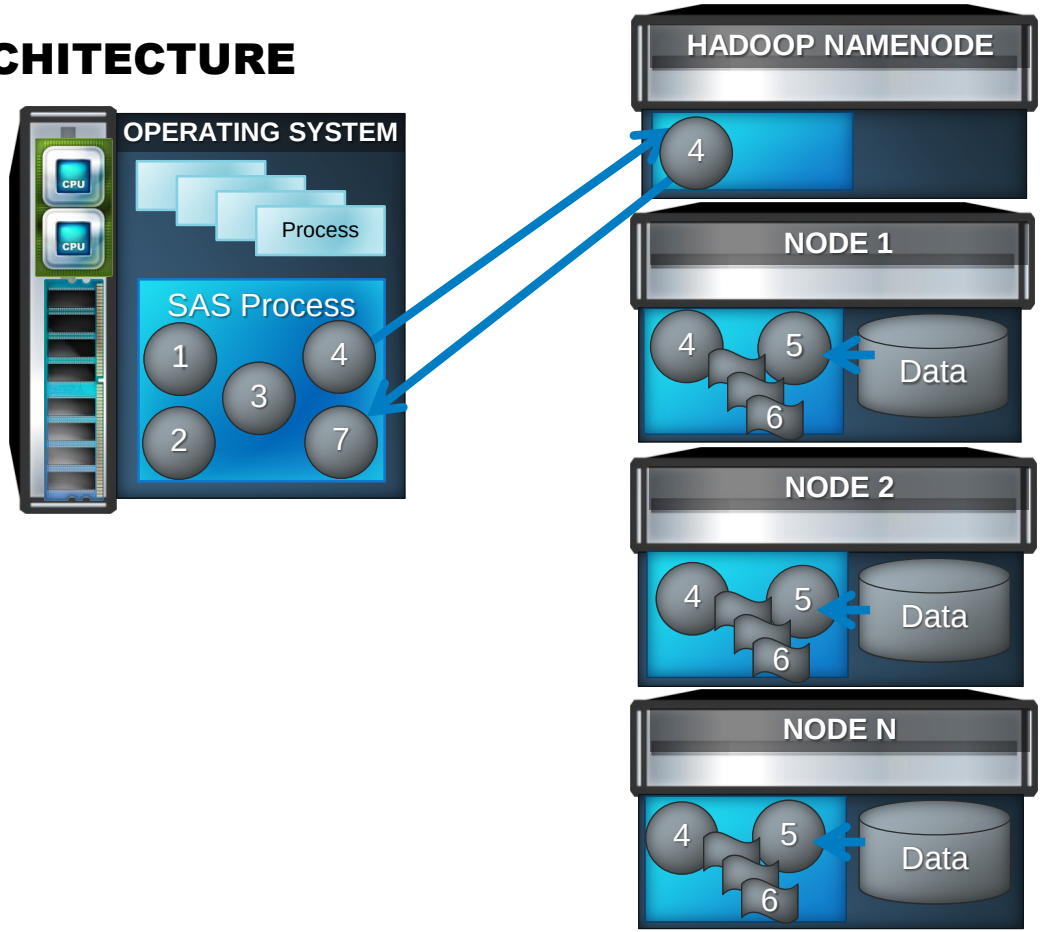
40 nodes x 96GB almost 4TB of memory

HPPROCS IN DISTRIBUTED ARCHITECTURE

```
libname a sashdat;
option set=gridhost="NAMENODE";
proc hpreg data=a.source;
   analytic stuff…
   performance nodes=all;
run;
```

SAS Process Steps:

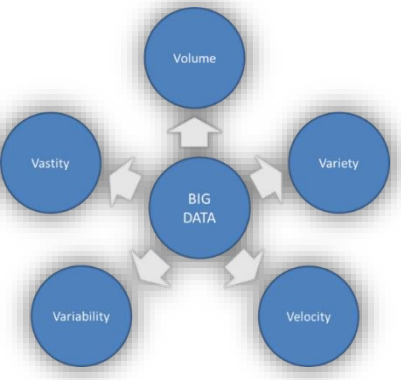
(1) SAS Process Starts on HW & O/S
(2) SAS sets up access library to disk
(3) SAS starts HPREG PROC
(4) Due to GRIDHOST and proper access engine setting, multi-threaded processes are started on grid nodes (via TKGrid)
(5) As TKGrid processes start up, ***ALL*** data is lifted into RAM from HDFS.
(6) Processing occurs in parallel against in memory data
(7) Results return to initiating process on SAS Server



ANALYTICAL CATEGORIES AND TARGET USAGE

Statistics	Data Mining	Text Mining	Forecasting	Econometrics	Optimization
• Binary target & continuous no. predictions • Linear, Non-Linear, & Mixed Linear modeling	• Complex relationships • Tree-based Classification • Variable Selection	• Parsing large-scale text collections • Extract entities • Auto. Stemming & synonym detection	• Large-scale, multiple hierarchy problems	• Probability of events • Severity of random events	• Local search optimization • Large-scale linear & mixed integer problems • Graph theory

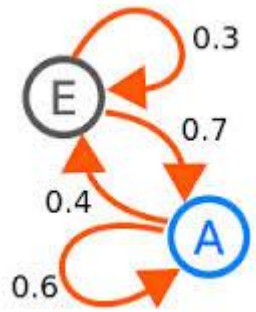
SAS ANALYTIC CATEGORIES



ANALYTICS STRATEGIES

- SAS collects many different methods:
  - many of them coming from traditional statistical inference analysis using SEMMA paradigm.
  - Many other coming from Bayesian inference
- Other coming from stochastic process analysis both for continue time and discrete events (discrete space).
- Time series forecasting: stochastic processes in continue time with continue space.
- Other coming from linear and not linear mixed models.
- Graph analysis

EVENTS PROCESSES

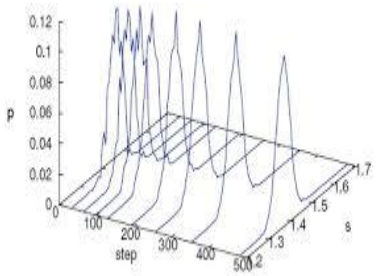


- When have a sequence of state changements (events) time depending we are managing stochastic process: or for continue time or discrete time SAS provides:
  - several approach to manage Markov chains considering also Bayesian posteriori prob distribution.

  Stochastic process, for all t, the conditional distribution of **Yt+1,given Y0,Y1,…,Yt** is identical to the conditional distribution of **Yt+1** given **Yt** alone. i.e, given **Yt, Yt+1** is conditional independent of **Y0,Y1,…,Yt−1.** So knowing the present state of a Markov chain,information about the past states does not help us **predict the future**
  P(Yt+1|Y0,Y1,…Yt)=P(Yt+1|Yt)

SAS ANALYTIC CATEGORIES

EVENTS PROCESSES

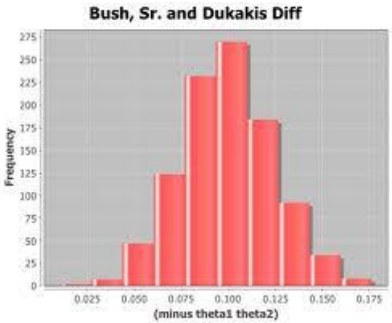


- **When have a sequence of state changements (events) time depending we are managing stochastic process: or for continue time or discrete time SAS provides:**

The Markov chain Monte Carlo (MCMC) method consists of a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. It combines the Monte Carlo method for sampling randomness and the Markov chain method for sampling independence with its stationary distribution.

SAS ANALYTIC CATEGORIES

EVENTS HISTORY



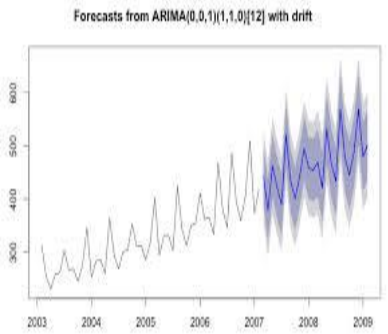
- **When have a sequence of state changements (events) time depending we are managing stochastic process: or for continue time or discrete time SAS provides:**

Event history is a range of methods scope defining time duration between an event variable change is status (level of event). These methods can be parameter based or not parameter based.

Events can be predicted also considering multinomial event probability to happen according several explanatory indicators.

SAS ANALYTIC CATEGORIES

TIME SERIES FORECAST

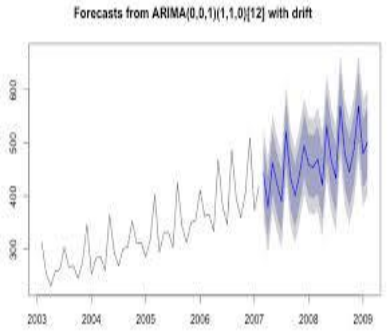


- **Time series forecasting: stochastic processes in discrete or continue time with continue space:**

  ARIMA analyzes and forecasts equally spaced univariate time series data, transfer function data, and intervention data by using the **autoregressive integrated moving-average** (ARIMA). An ARIMA model predicts a value in a response time series **as a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series**.

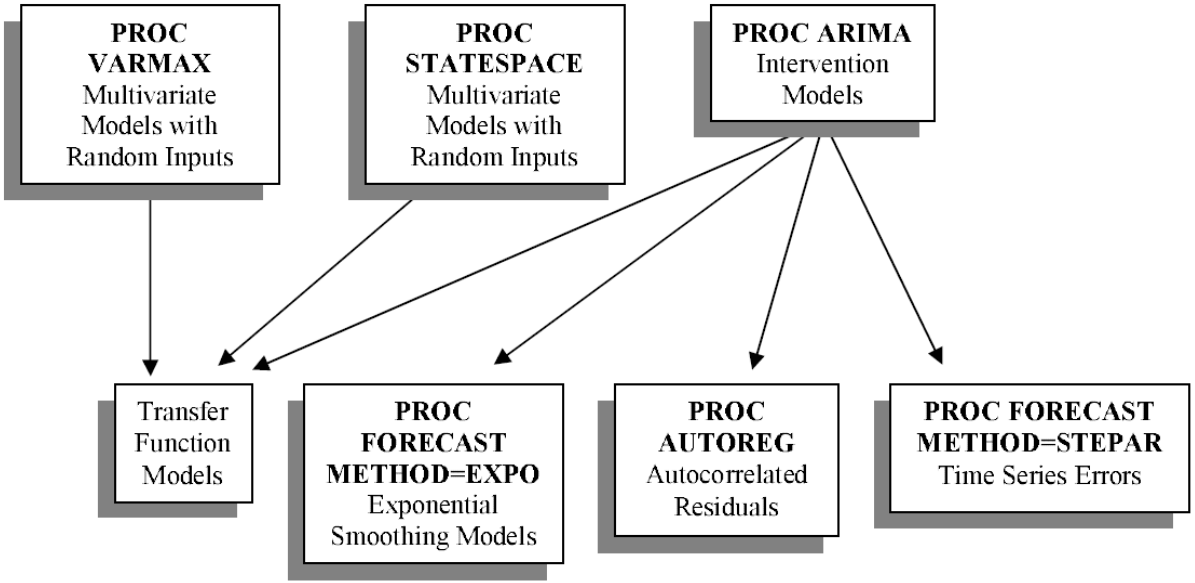
  analyzes and forecasts equally spaced univariate time series data by using an **unobserved components model (UCM)** also called *structural models* in the time series literature

SAS ANALYTIC CATEGORIES

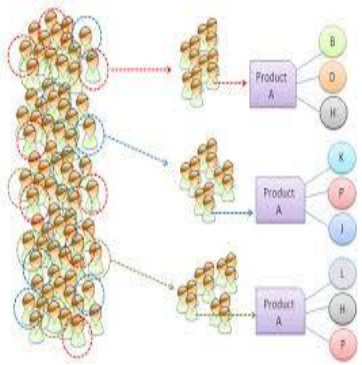


TIME SERIES FORECAST

- Time series forecasting: stochastic processes in continue time with continue space.



RECOMMENDATION SYSTEMS

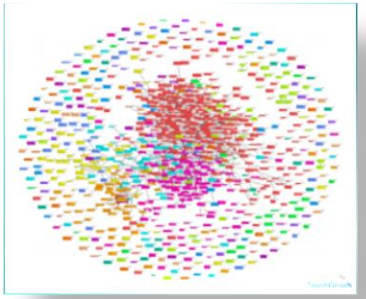


There is an extensive class of Web applications that involve predicting **user responses to options**. Such a facility is called a *recommendation system*.
Recommendation systems use a number of different methods.

Content-based systems: methods are based on a description of the item and a profile of the user's preference. keywords are used to describe the items beside, a user profile is built to indicate the type of item this user likes.

Collaborative filtering: Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users.

GRAPH ANALYSIS



Graph measures gives metric about relationships among nodes links among them into a net:
There are 2 approach a graph measures:

- **Random**
- **Structural**

Several measures characterize it:

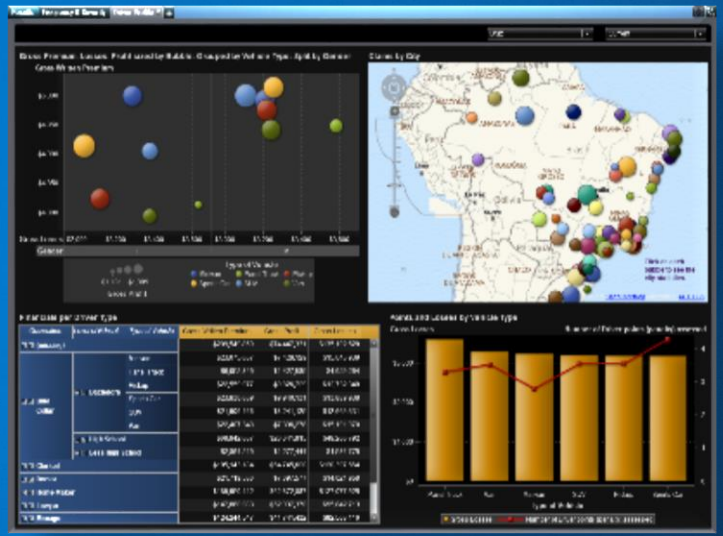
STRENGTH: related to the relationship with the 'neighborhood‘

POSITION: linked to the "paths" in the subnet

PRESTIGE: related to the interaction with the whole subnet

SAS® VISUAL ANALYTICS

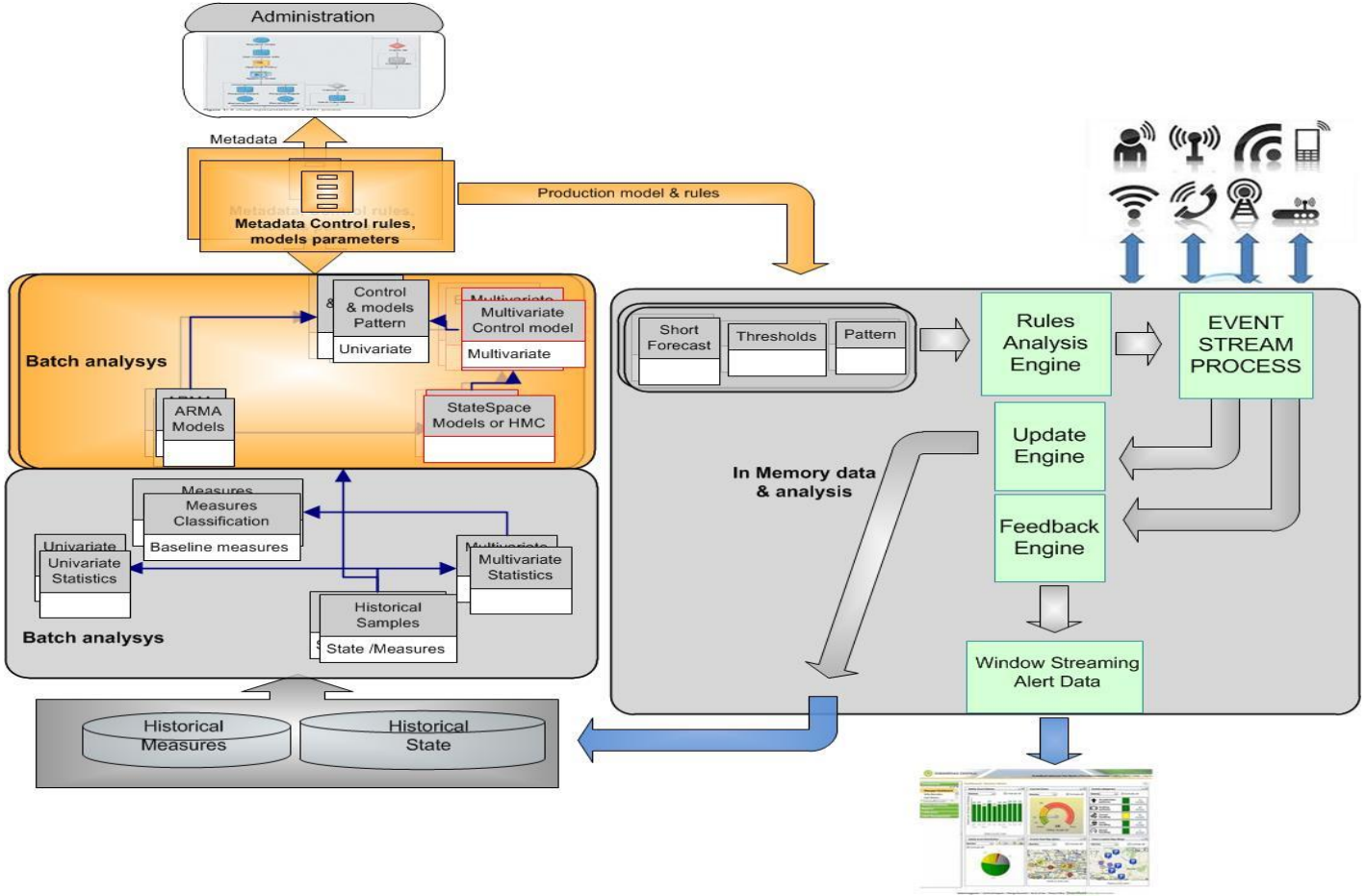
VISUALIZATION DRIVEN BY ANALYTICS



EXPLORATION AND VISUALIZATION POWER OF ANALYTICS: USING UNIVARIATE AND MULTIVARIATE GRAPHICAL STATISTICS

sas THE POWER TO KNOW.

EXAMPLE ARCHITECTURE: REAL TIME MONITORING



QUESTIONS