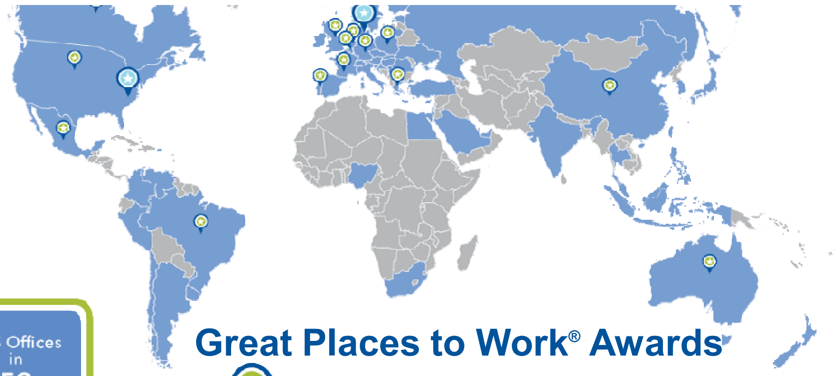
SAS ANALYTIC SOLUTIONS RUNNING ON A HADOOP CLUSTER USING YARN

JAMES KOCHUBA

MARKET LEADER IN DATA & ANALYTICS



SAS Offices in 59 Countries

Great Places to Work® Awards

15 COUNTRIES

2 MULTINATIONAL

SAS 23%

Industry Average 16%

Revenue Reinvested in R&D

3,400

SAS® Visual Analytics Customer Sites

24%

SAS® Cloud Analytics Revenue Growth

SAS® - Hadoop visualization and analytics solutions

#1 PREDICTIVE ANALYTICS
ADVANCED ANALYTICS

As Ranked by IDC

SAS customers represent **90%** of Fortune Global 500® companies

3+ Billion 2014 REVENUE

Customers in 139 countries at 70,000 sites

35% MARKETSHARE

3 DECADES OF EXPERIENCE

SAS Background

Millions of analytical procedures running at **65,000 sites**

Analytics applied to thousands of business issues

41,000 customers in 135 countries

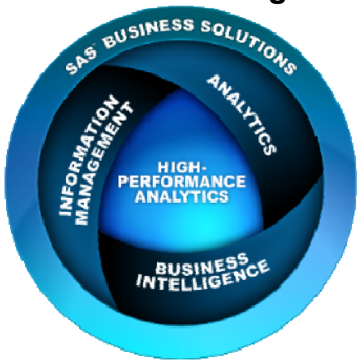
Three-plus decades of experience

$650 million annually in advanced analytics revenue

Total Yearly Revenues $2.8B

IDC ranks **SAS No. 1 in advanced analytics** with a market share of **36.2%**

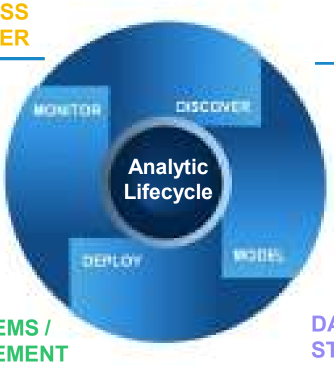
SAS Core Technologies



SAS Advanced Analytics

- Statistics
- Predictive Modeling
- Data Mining
- Text analytics
- Forecasting & Econometrics
- Quality Improvement
- Operations Research
- Data Visualization
- Model Management and Deployment

SAS and the Analytic Lifecycle



BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation
Author Rule Logic

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation

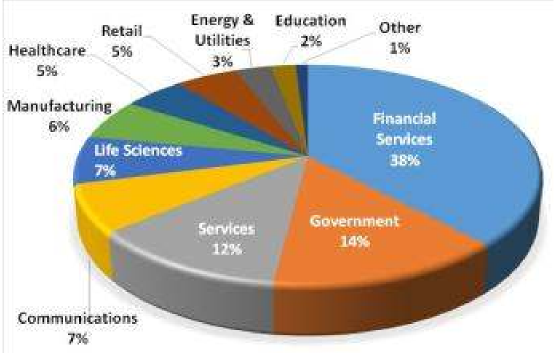
DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

Solution Lines

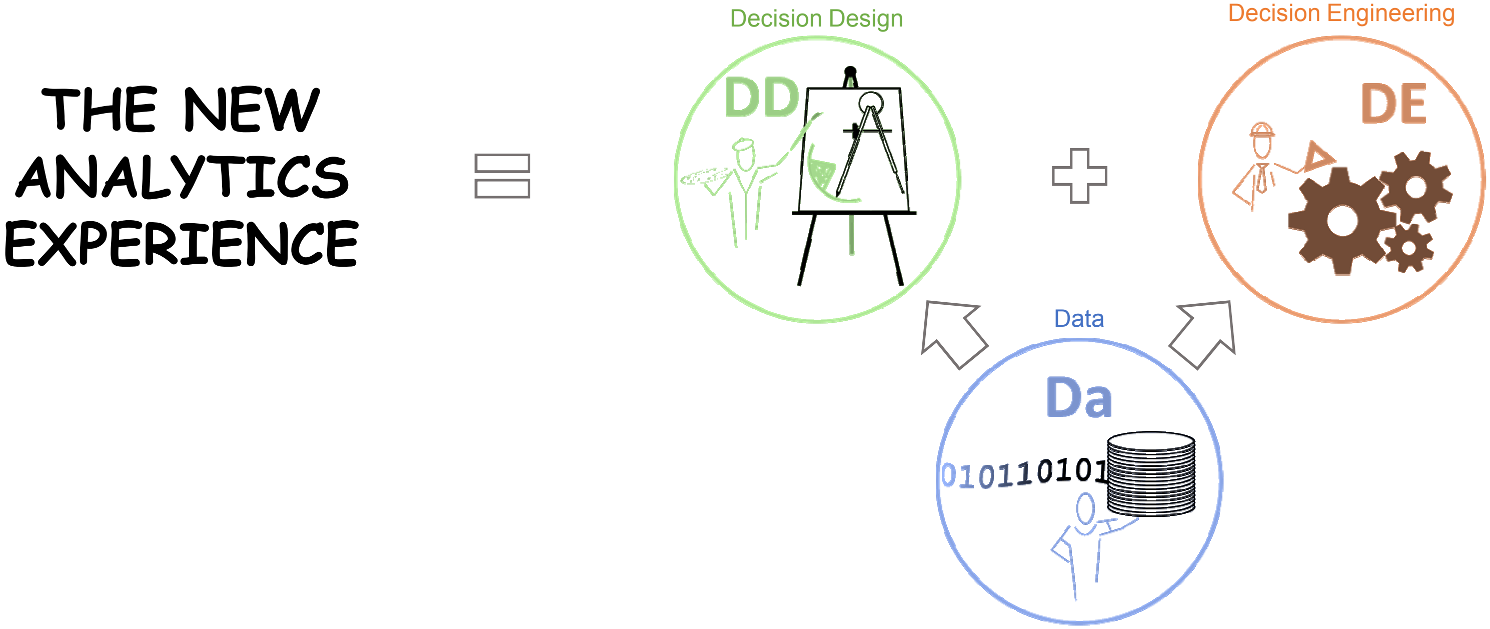
- Analytics
- Business Intelligence
- Customer Intelligence
- Financial Intelligence
- Foundation Tools
- Fraud & Security Intelligence
- Governance, Risk & Compliance
- High-Performance Analytics
- Information Management
- IT & CIO Enablement
- OnDemand Solutions
- Performance Management
- Risk Management
- Supply Chain Intelligence
- Sustainability Management

Industries



KICKOFF 2015 | THE NEW ANALYTICS *EXPERIENCE*

- SAS is uniquely positioned to :
  - Enable and Empower the new Analytics Culture;
  - BRIDGE the gaps between **Decision Design**, **Decision Engineering**, and the **Data**.



KICKOFF 2015 | **THE NEW ANALYTICS EXPERIENCE**

The “ *Art* ” | The “ Process ”

DECISION DESIGN	**DECISION ENGINEERING**
Data is a Raw Material	**Data** is a finished product
Flexible, ad hoc	Established, documented process
Prototyping	Governance (over data, process, technology)
Data Scientists, Analysts, Smart Creatives	Engineers, DBA, IT
Open Source, “whatever works”	Approved architecture
Departmental, personal	Enterprise
Innovative, Experimental, Groundbreaking	***Productionized, Scalable, Repeatable***
DATA	
No amount or complexity is unsurmountable	

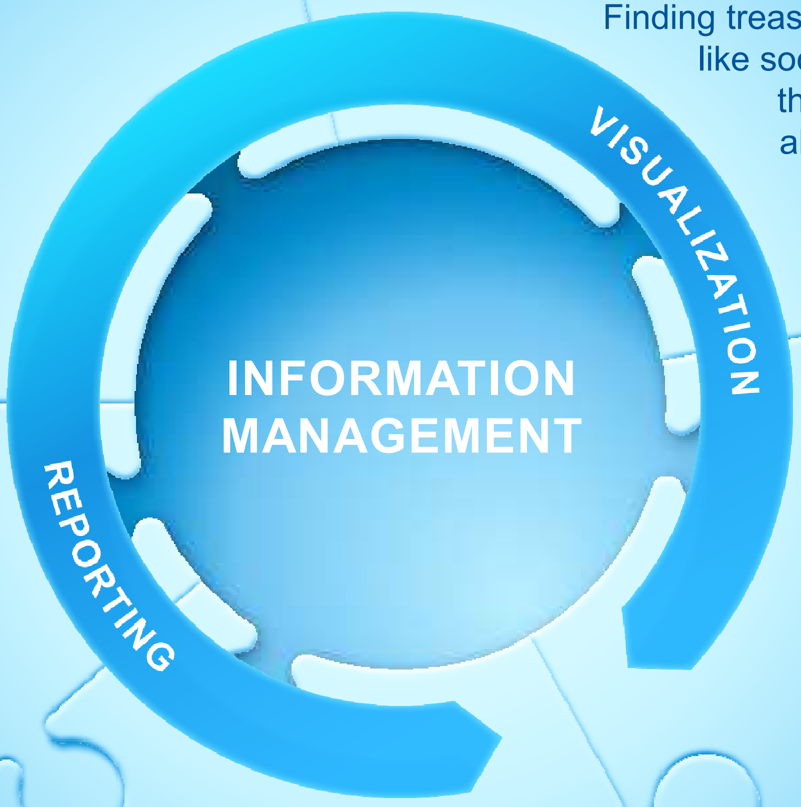
ANALYTICS

FORECASTING

Leveraging historical data to drive better insight into decision-making for the future

DATA MINING

Mine transaction databases for data of spending patterns that indicate a stolen card



INFORMATION MANAGEMENT

VISUALIZATION

REPORTING

STATISTICS

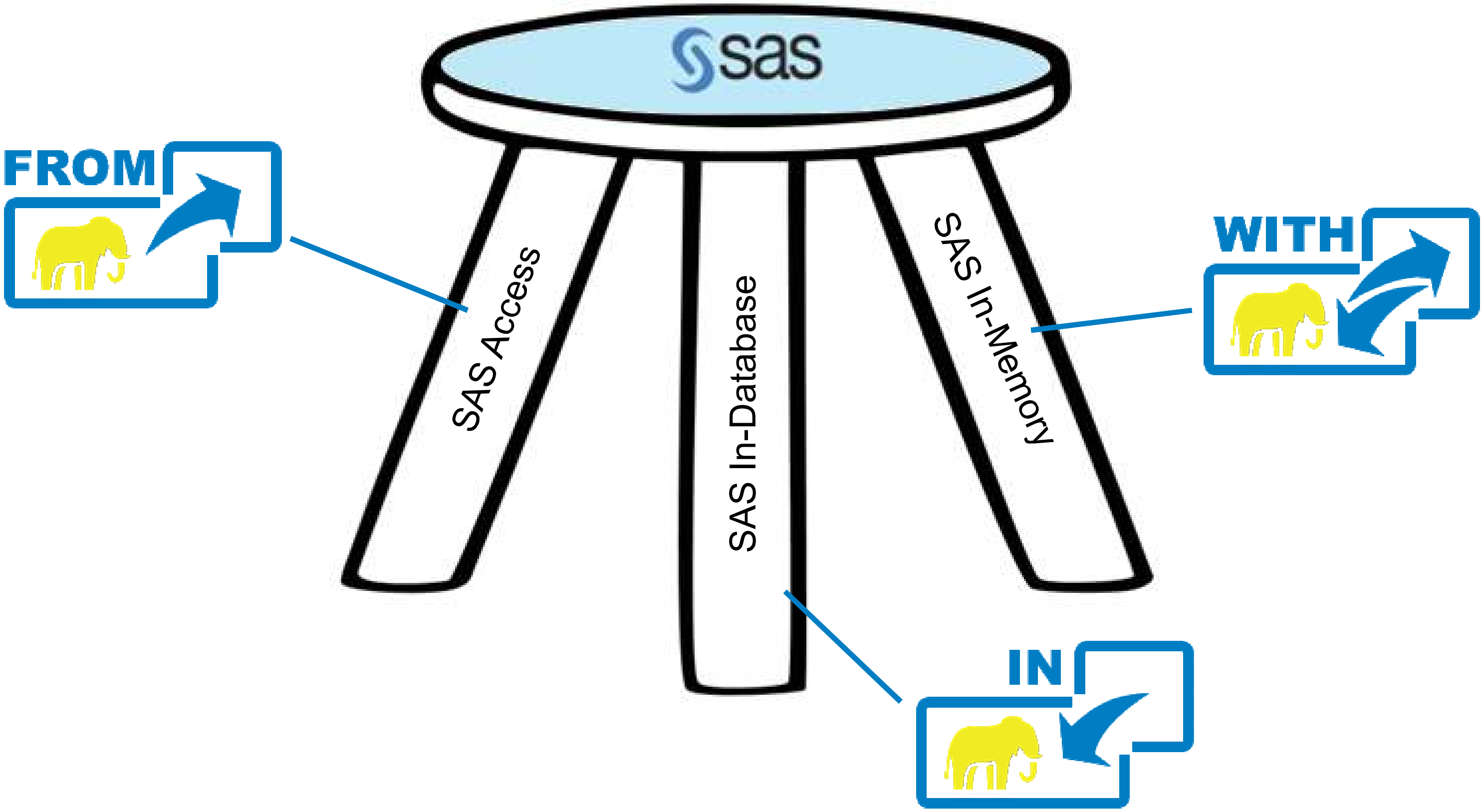
TEXT ANALYTICS

Finding treasures in unstructured data like social media or survey tools that could uncover insights about consumer sentiment

OPTIMIZATION

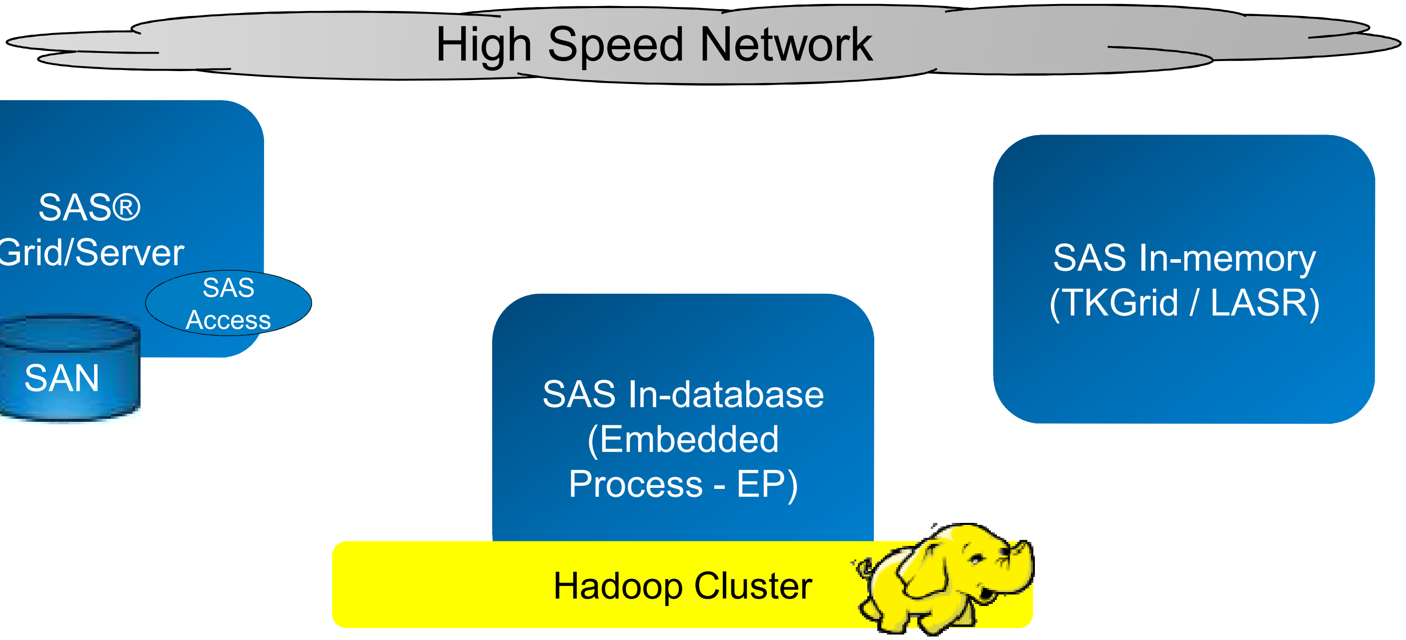
Analyze massive amounts of data in order to accurately identify areas likely to produce the most profitable results

SAS | CRITICAL SAS COMPONENTS FOR HADOOP



ARCHITECTURE REVIEW

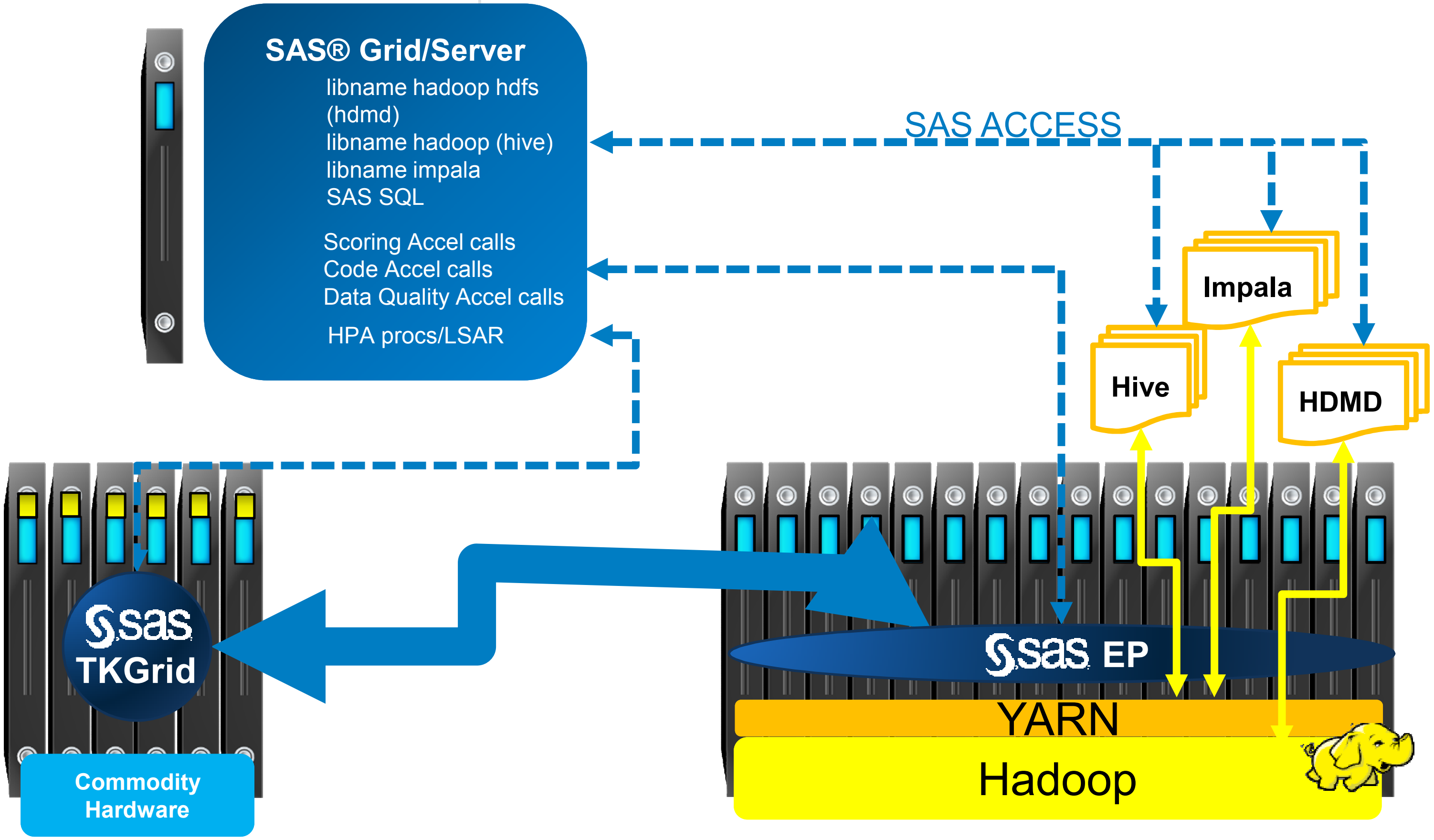
SAS SOFTWARE WITH HADOOP



In-memory

- Private
- Public

SAS AND HADOOP | TRADITIONAL SAS WITH HADOOP



SAS® Grid/Server

libname hadoop hdfs (hdmd)
libname hadoop (hive)
libname impala
SAS SQL

Scoring Accel calls
Code Accel calls
Data Quality Accel calls

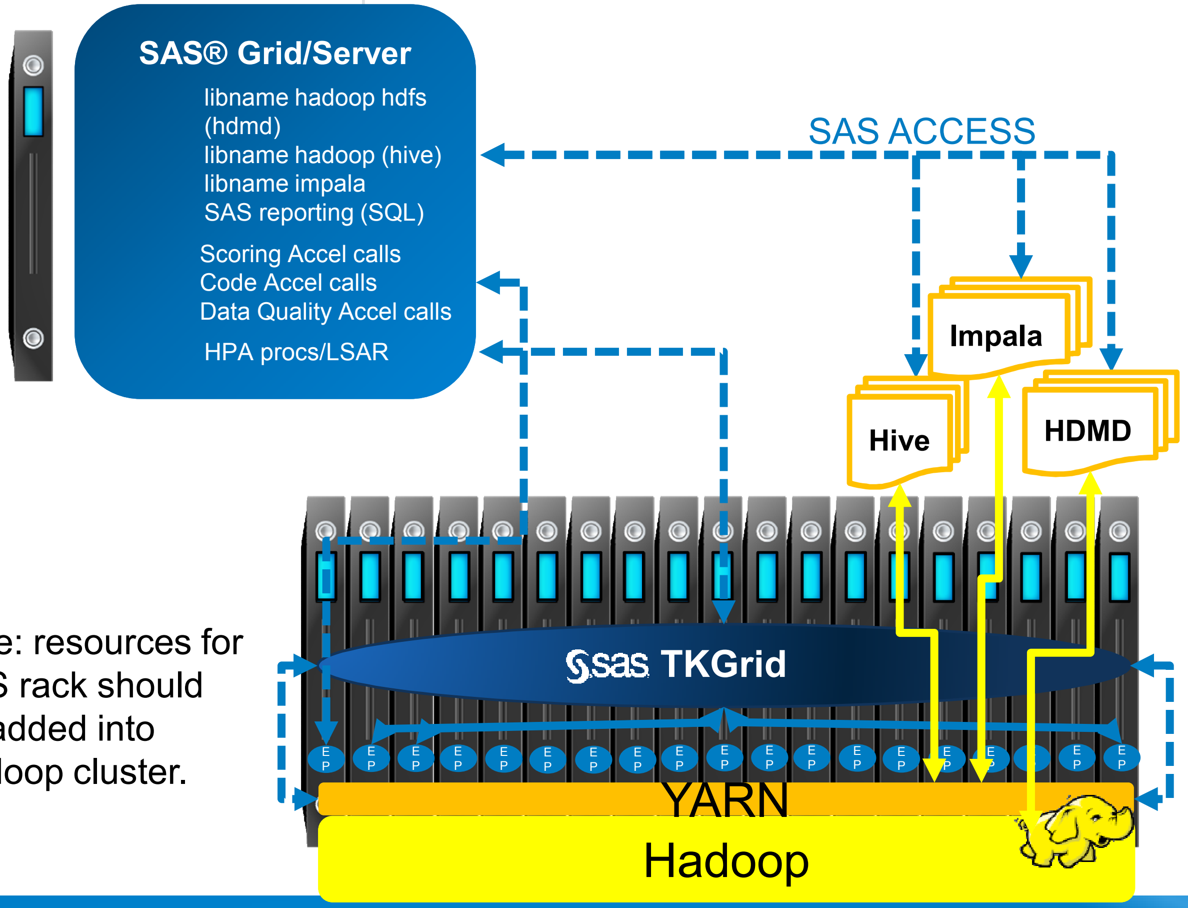
HPA procs/LSAR

Yarn is effecting:

- SAS Hive and Impala calls
- SAS EP (Mapreduce)

No yarn effect on HDMD since that goes directly to HDFS

SAS AND HADOOP | SAS ON HADOOP



SAS® Grid/Server

libname hadoop hdfs (hdmd)
libname hadoop (hive)
libname impala
SAS reporting (SQL)

Scoring Accel calls
Code Accel calls
Data Quality Accel calls

HPA procs/LSAR

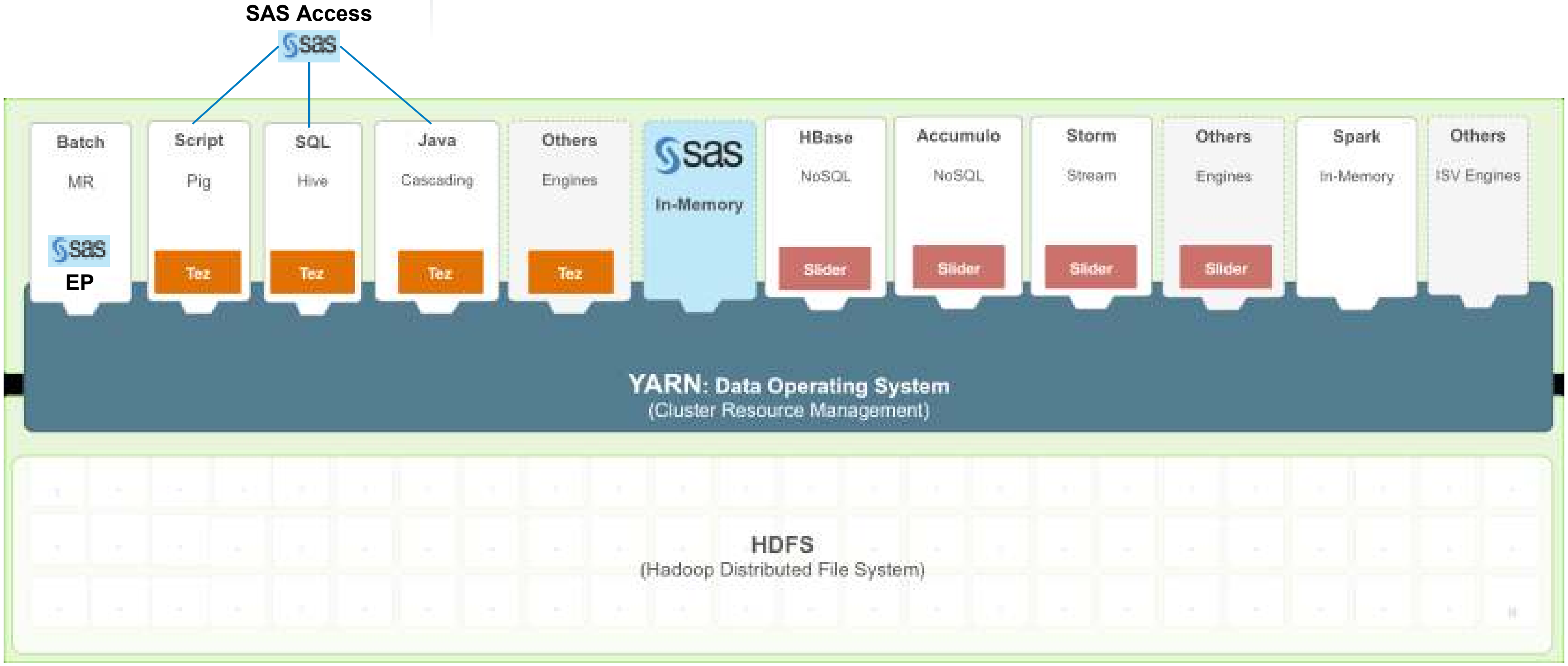
Note: resources for SAS rack should be added into Hadoop cluster.

Yarn is effecting:

- SAS Hive calls
- SAS EP (mapreduce)
- **SAS In-memory**
  - To start TKGrid process, we will work with YARN

No yarn effect on HDMD since that goes directly to HDFS

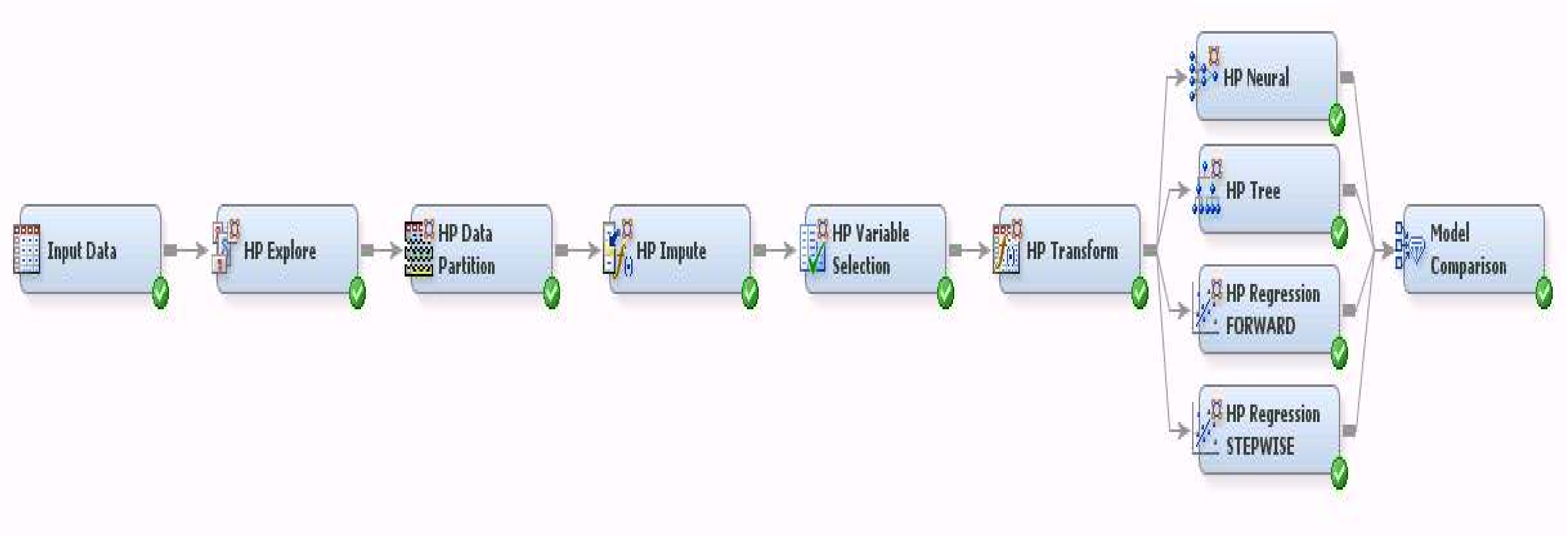
SAS AND YARN | WHERE DOES SAS FIT IN?



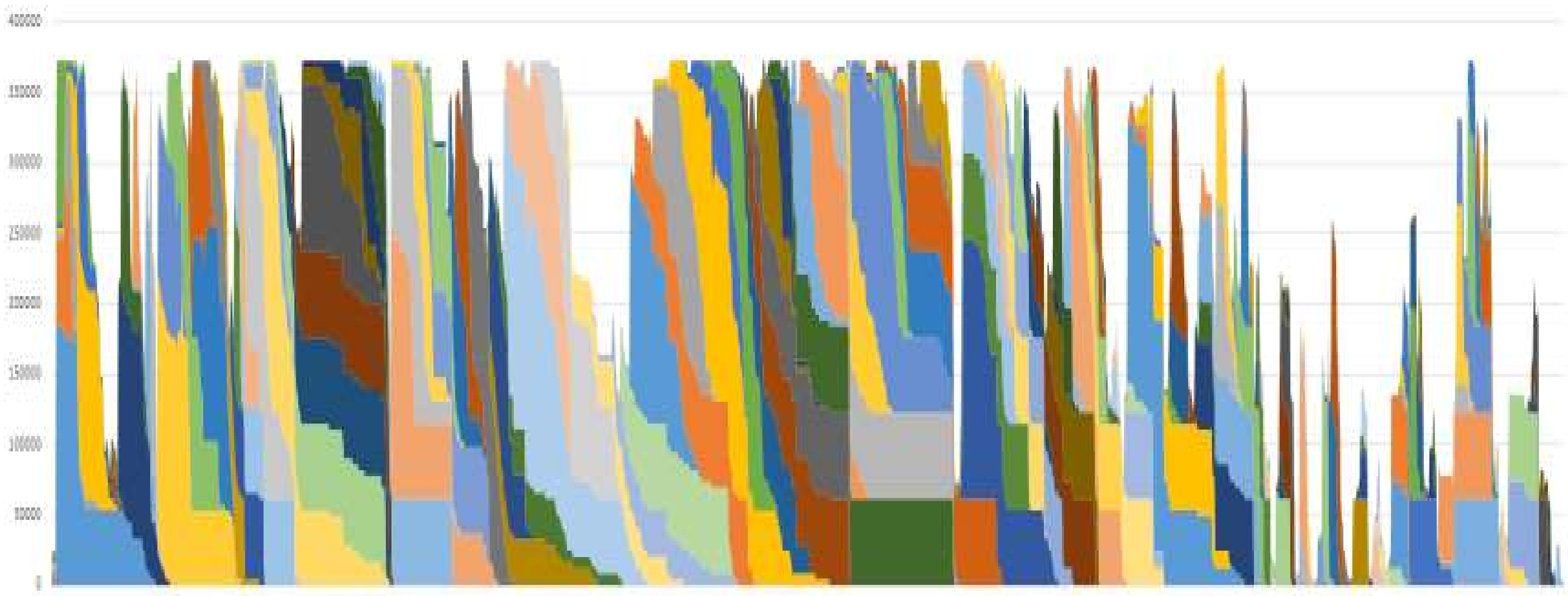
*Picture Created by Arun Murthy - Hortonworks

http://blogs.sas.com/content/datamanagement/2014/08/20/sas-high-performance-capabilities-with-hadoop-yarn/

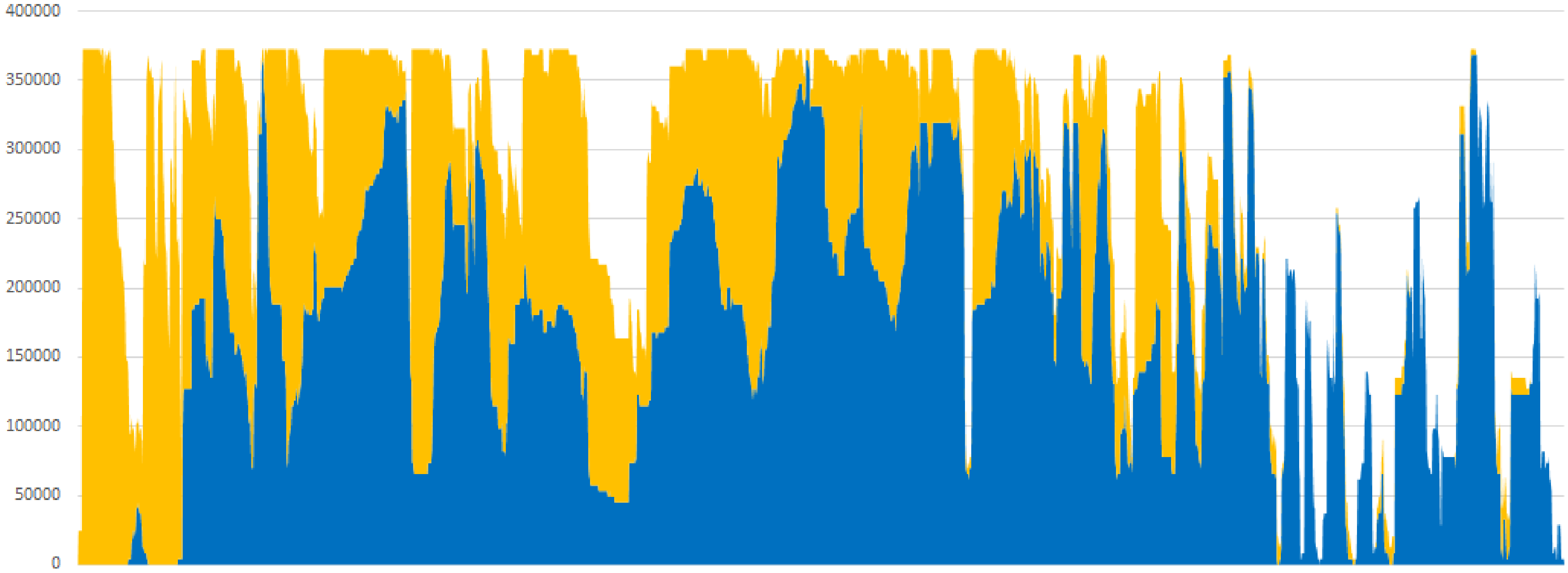
SAS MODEL | SAS HPDM EXAMPLE



YARN VIEW | SHARED SAS AND HADOOP ENVIRONMENT

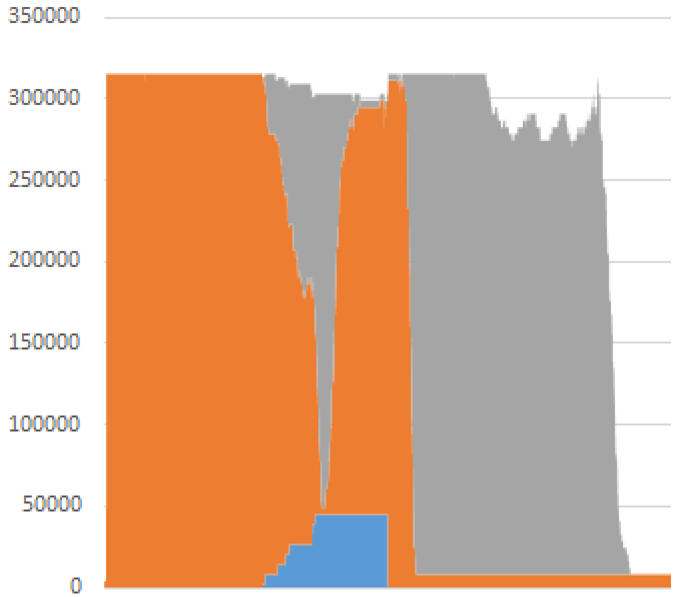


YARN VIEW | SAS VS OTHER WORK

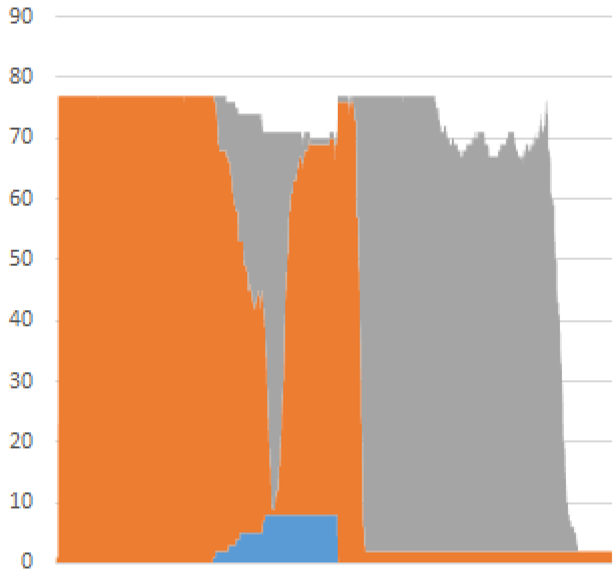


YARN VIEW | SMALL SAS APPLICATION WITH BACKGROUND

Memory Usage



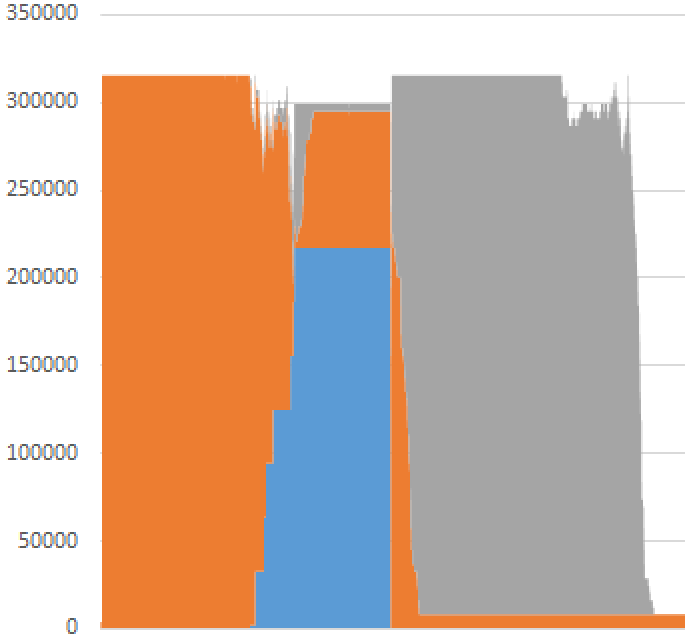
Number of Container



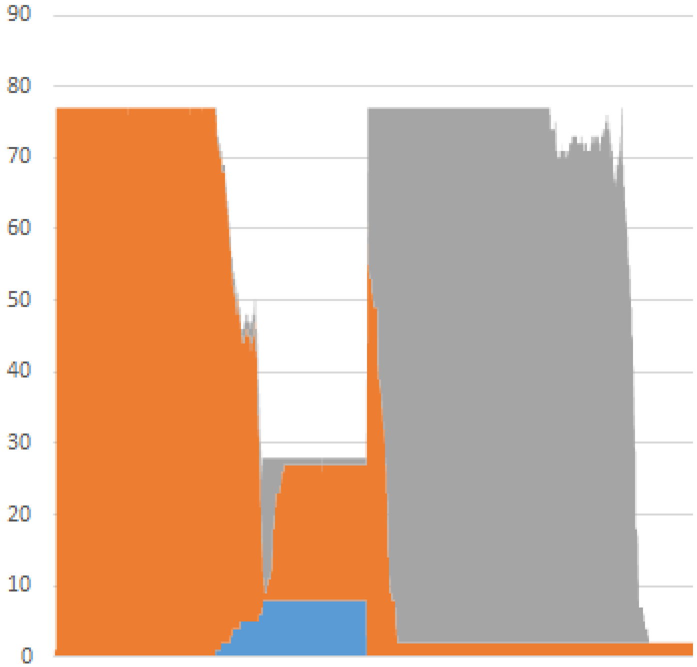
Compress data ~31 GB (20,000,000 observations, 50 variables)

YARN VIEW | LARGER SAS APPLICATION WITH BACKGROUND

Memory Usage

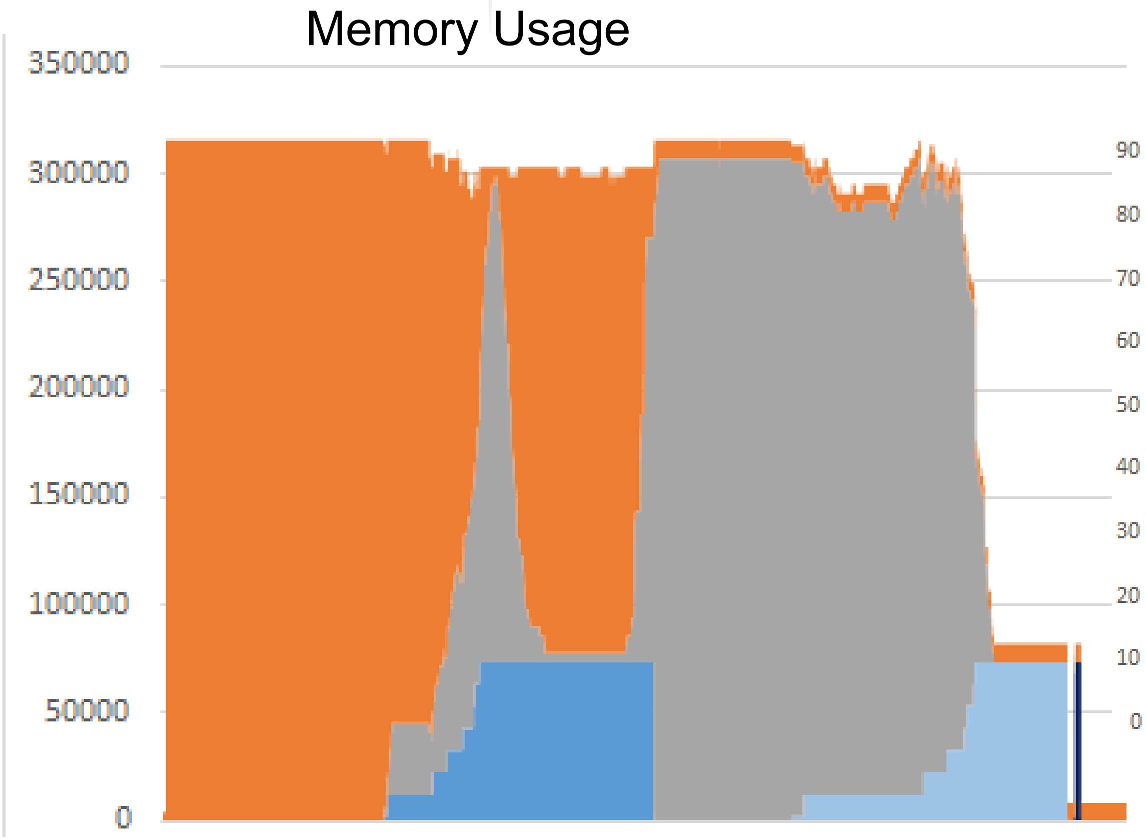


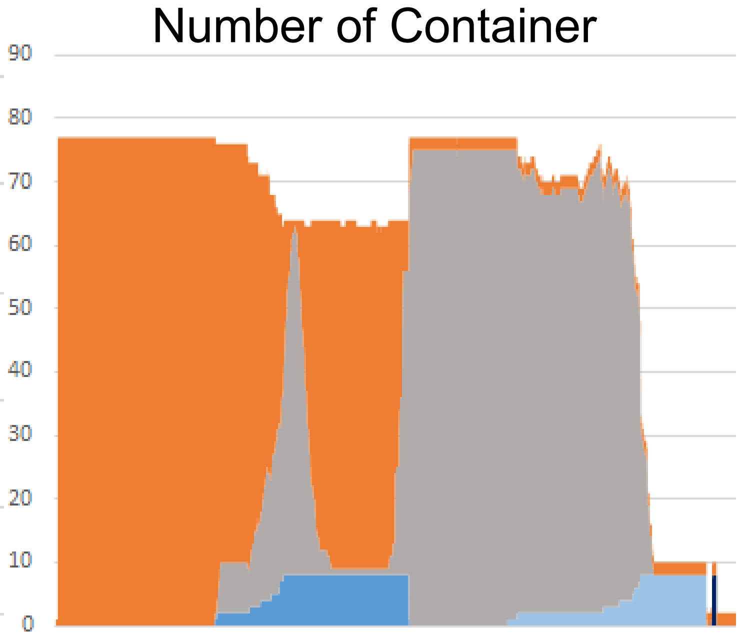
Number of Container



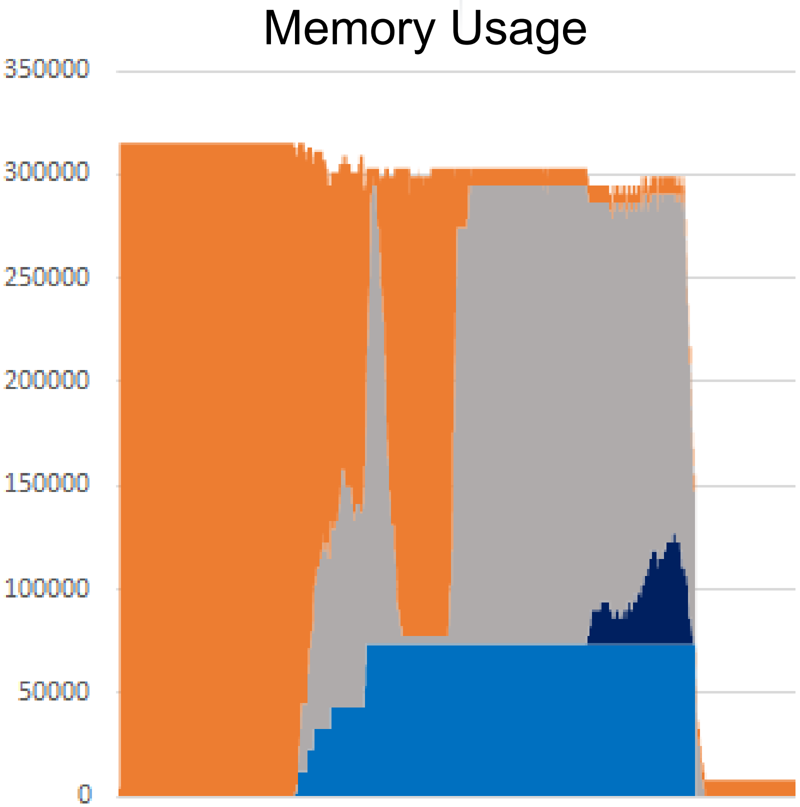
Compress data ~183 GB (120,000,000 observations, 50 variables)

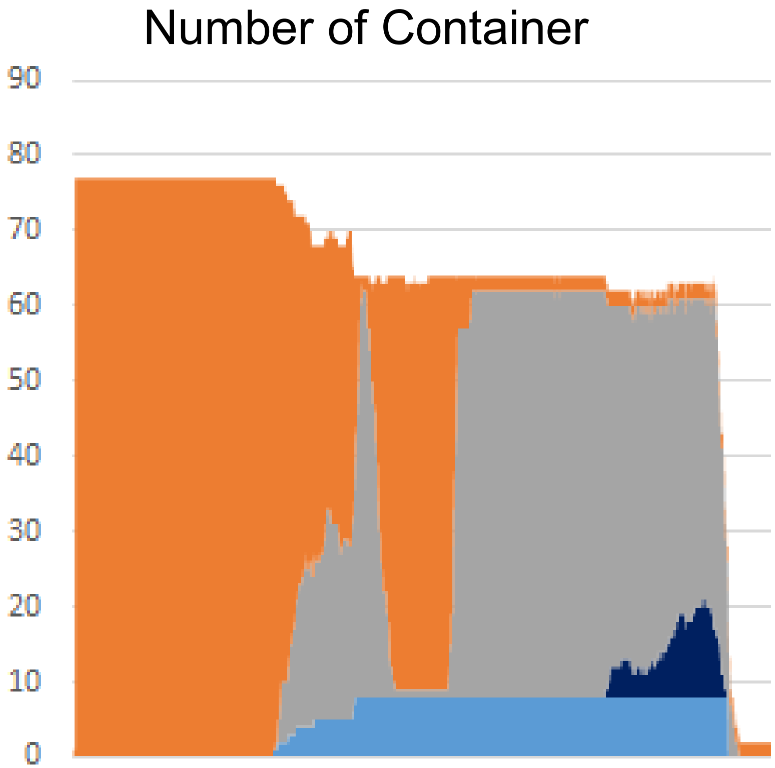
YARN VIEW | SIMPLE SAS MODEL WITH BACKGROUND





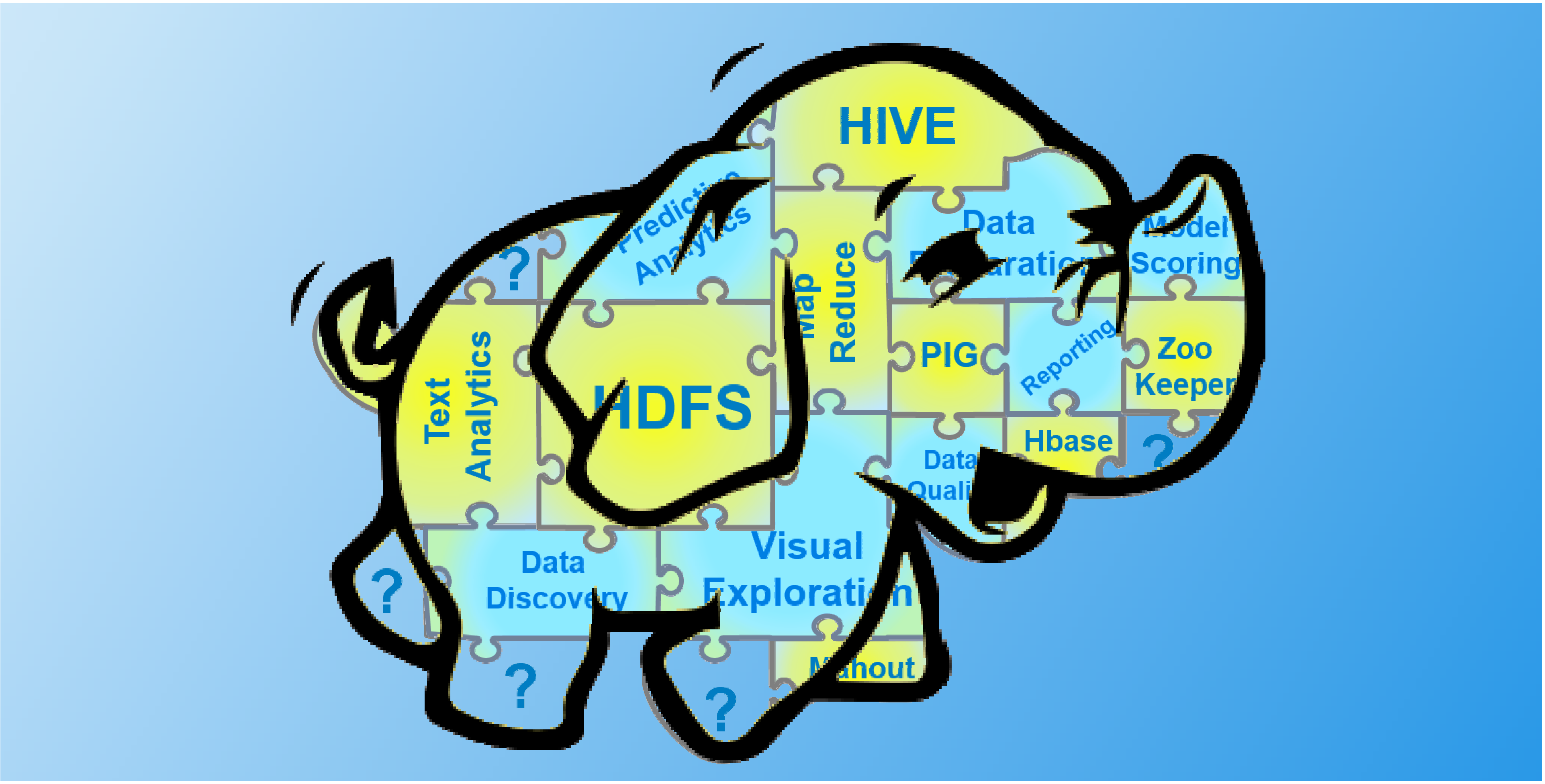
YARN VIEW | SAS APPLICATION LOADING HIVE WITH BACKGROUND





CLIENT ISSUES | LESSON LEARNS

- Minimum container memory size can product wasted memory resources
  - MapReduce application does not use all memory
  - Smaller applications pushed into large containers (application master to simple applications)
- MapReduce tuning
- Dependency jobs require queue to help
  - SAS In-memory using SAS EP to lift data into memory
- Queue happy craves up cluster too much
- Monitor real resource usage vs containers
  - Focus on application tuning
- SAS YARN workshop


HIVE
Predictive Analytics
Map Reduce
Data Preparation
Model Scoring
Text Analytics
HDFS
PIG
Reporting
Zoo Keeper
Hbase
Data Quality
Data Discovery
Visual Exploration
Mahout
?
?
?
?
?
?

Free Software Trial!

Leading provider for:

- Data **Preparation**
- Data **Visualization**
- Data **Analysis**

sas.com/strata-trial

Enter for a chance to win a GoPro HERO4!

Booth
1022

QUESTIONS

THE POWER TO KNOW.

YARN TUNING | BASIC YARN SETTINGS

Property Name	Description
yarn.nodemanager.resource.memory-mb	Amount of physical memory, in MiB, that can be allocated for containers.
yarn.scheduler.minimum-allocation-mb	The minimum allocation for every container request at the RM, in MBs. Memory requests lower than this won't take effect, and the specified value will get allocated at minimum.
yarn.scheduler.maximum-allocation-mb	Largest Container allowed. A Multiple of the minimum-allocation-mb above Depending on your setup you may want to allow the entire node for MR, or restrict it to smaller then a node to prevent potential malicious actions.
yarn.nodemanager.resource.cpu-vcores	Number of virtual CPU cores that can be allocated for containers. This value covers all applications and their containers running on this node and or physical system.
yarn.scheduler.minimum-allocation-vcores	The smallest number of virtual CPU cores that can be requested per container.
yarn.scheduler.maximum-allocation-vcores	The largest number of virtual CPU cores that can be requested per container.
yarn.resourcemanager.scheduler.class	The class used for resource manager (note Hortonworks and Cloudera used different defaults and today, they do prompt writing custom classes)

YARN TUNING | MAPREDUCE SETTINGS

Property Name	Description
mapreduce.map.memory.mb	The size of the container for the Mapper task
mapreduce.map.java.opts	The java opts for the Mapper JVM, make sure that the max heap is less then the size of the container.
mapreduce.reduce.memory.mb	The size of the container for the Reducer task
mapreduce.reduce.java.opts	The java opts for the Reducer JVM, make sure that the max heap is less then the size of the container.
mapreduce.job.reduce.slowstart.completedmaps	Fraction of the number of maps in the job which should be complete before reduces are scheduled for the job.

YARN TUNING | QUEUES

- Scheduler queuing
  - FairScheduler -
  - CapacityScheduler – queues
- Cloudera queuing
  - Dynamic Pools