SAS ANALYTIC SOLUTIONS RUNNING ON A HADOOP CLUSTER USING YARN
JAMES KOCHUBA
MARKET LEADER IN DATA & ANALYTICS

Great Places to Work® Awards
- 15 COUNTRIES
- 2 MULTINATIONAL

SAS® Visual Analytics Customer Sites
- 3,400

SAS® Cloud Analytics Revenue Growth
- 24%

SAS® - Hadoop visualization and analytics solutions

SAS Industry Average Revenue Reinvested in R&D
- 23%
- 16%

#1 PREDICTIVE ANALYTICS ADVANCED ANALYTICS
As Ranked by IDC
3+ Billion
2014 REVENUE
Customers in 139 countries at 70,000 sites
35% MARKETSHARE
3 DECADES OF EXPERIENCE

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

- Analytic Lifecycle
  - Domain Expert Makes Decisions
  - Evaluates Processes and ROI
  - Model Validation
  - Model Deployment
  - Model Monitoring
  - Data Preparation

- High-Performance 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

**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**

- Domain Expert
  - Exploratory Analysis
  - Descriptive Segmentation
  - Predictive Modeling

- Data Exploration
  - Exploratory Analysis
  - Descriptive Segmentation
  - Predictive Modeling

- Data Visualization
  - Exploratory Analysis
  - Descriptive Segmentation
  - Predictive Modeling

- Business Manager
  - Model Validation
  - Model Deployment
  - Model Monitoring
  - Data Preparation

- Business Analyst
  - Model Validation
  - Model Deployment
  - Model Monitoring
  - Data Preparation

- Data Miner / Statistician
  - Model Validation
  - Model Deployment
  - Model Monitoring
  - Data Preparation

**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**

- Financial Services: 38%
- Manufacturing: 5%
- Retail: 3%
- Government: 14%
- Life Sciences: 12%
- Energy & Utilities: 9%
- Communications: 7%
- Healthcare: 5%
- Services: 5%
- Other: 5%
- Education: 2%
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.
### The “Art” vs. The “Process”

<table>
<thead>
<tr>
<th>DECISION DESIGN</th>
<th>DECISION ENGINEERING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data is a Raw Material</td>
<td>Data is a finished product</td>
</tr>
<tr>
<td>Flexible, ad hoc</td>
<td>Established, documented process</td>
</tr>
<tr>
<td>Prototyping</td>
<td>Governance (over data, process, technology)</td>
</tr>
<tr>
<td>Data Scientists, Analysts, Smart Creatives</td>
<td>Engineers, DBA, IT</td>
</tr>
<tr>
<td>Open Source, “whatever works”</td>
<td>Approved architecture</td>
</tr>
<tr>
<td>Departmental, personal</td>
<td>Enterprise</td>
</tr>
</tbody>
</table>

**Innovative, Experimental, Groundbreaking**

**Productionized, Scalable, Repeatable**

**DATA**

*No amount or complexity is unsurmountable*
Finding treasures in unstructured data like social media or survey tools that could uncover insights about consumer sentiment.

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.

TEXT ANALYTICS

INFORMATION MANAGEMENT

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

STATISTICS

REPORTING

VISUALIZATION
SAS CRITICAL SAS COMPONENTS FOR HADOOP

FROM

SAS Access

WITH

SAS In-Memory

IN

SAS In-Database
ARCHITECTURE REVIEW

SAS SOFTWARE WITH HADOOP

High Speed Network

- SAS® Grid/Server
  - SAN
  - SAS Access

- SAS In-database (Embedded Process - EP)

- SAS In-memory (TKGrid / LASR)
  - In-memory
    - Private
    - Public

- Hadoop Cluster
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

SAS ACCESS

Yarn is effecting:
• SAS Hive and Impala calls
• SAS EP (Mapreduce)
No yarn effect on HDMD since that goes directly to HDFS

TKGrid
Commodity Hardware

Hadoop
YARN

SAS ACCESS

Impala

Hive
HDMD

EP
SAS AND 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

SAS ACCESS

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

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

Hadoop

TKGrid

Hive

Impala

HDMD

YARN

Scoring Accel calls
Code Accel calls
Data Quality Accel calls
HPA procs/LSAR

No yarn effect on HDMD since that goes directly to HDFS
**SAS AND YARN**

WHERE DOES SAS FIT IN?

*SAS Access*

---

**YARN: Data Operating System**
(Cluster Resource Management)

---

**HDFS**
(Hadoop Distributed File System)

---

*Picture Created by Arun Murthy - Hortonworks*

---

YARN VIEW

SHARED SAS AND HADOOP ENVIRONMENT
Compress data ~31 GB (20,000,000 observations, 50 variables)
Compress data ~183 GB (120,000,000 observations, 50 variables)
YARN VIEW
SAS APPLICATION LOADING HIVE WITH BACKGROUND

Memory Usage

Number of Container
CLIENT ISSUES  LESSON LEARNS

• Minimum container memory size can produce 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
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
### YARN TUNING

#### BASIC YARN SETTINGS

<table>
<thead>
<tr>
<th>Property Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>yarn.nodemanager.resource.memory-mb</td>
<td>Amount of physical memory, in MiB, that can be allocated for containers.</td>
</tr>
<tr>
<td>yarn.scheduler.minimum-allocation-mb</td>
<td>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.</td>
</tr>
<tr>
<td>yarn.scheduler.maximum-allocation-mb</td>
<td>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.</td>
</tr>
<tr>
<td>yarn.nodemanager.resource.cpu-vcores</td>
<td>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.</td>
</tr>
<tr>
<td>yarn.scheduler.minimum-allocation-vcores</td>
<td>The smallest number of virtual CPU cores that can be requested per container.</td>
</tr>
<tr>
<td>yarn.scheduler.maximum-allocation-vcores</td>
<td>The largest number of virtual CPU cores that can be requested per container.</td>
</tr>
<tr>
<td>yarn.resourcemanager.scheduler.class</td>
<td>The class used for resource manager (note Hortonworks and Cloudera used different defaults and today, they do prompt writing custom classes)</td>
</tr>
</tbody>
</table>
### YARN TUNING  
#### MAPREDUCE SETTINGS

<table>
<thead>
<tr>
<th>Property Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mapreduce.map.memory.mb</td>
<td>The size of the container for the Mapper task</td>
</tr>
<tr>
<td>mapreduce.map.java.opts</td>
<td>The java opts for the Mapper JVM, make sure that the max heap is less than the size of the container.</td>
</tr>
<tr>
<td>mapreduce.reduce.memory.mb</td>
<td>The size of the container for the Reducer task</td>
</tr>
<tr>
<td>mapreduce.reduce.java.opts</td>
<td>The java opts for the Reducer JVM, make sure that the max heap is less than the size of the container.</td>
</tr>
<tr>
<td>mapreduce.job.reduce.slowstart.completedmaps</td>
<td>Fraction of the number of maps in the job which should be complete before reduces are scheduled for the job.</td>
</tr>
</tbody>
</table>
YARN TUNING

| QUEUES |

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