HVORDAN KOMME I GANG MED SAS OG HADOOP?

JONAS LIE NIELSEN, PRINCIPAL SOLUTIONS ARCHITECT, STRATEGY, SAS INSTITUTE
We think this will provide some major insight...

What is it?

Not sure. But it’s BIG. Don’t we have SAS on that badoop thing?

It’s called Hadoop.

Whatever.

I’ll take a look...
ABOUT THE SURVEY

The Nordic countries is expected to be more than 1 year behind US and UK in Big Data Analytics and Hadoop adoption. We wanted to find out more:

- Understand Nordic adoption and maturity by country and industry
- What is the primary reasons, use cases and obstacles
- Understand where to sell the value of SAS – as we don’t sell Hadoop!

More than 300 Nordic companies responded through phone (Norstat) and online questionnaire (SAS Institute).

See [www.nordichadoopsurvey.com](http://www.nordichadoopsurvey.com) for results
“SAS, a known leader in enterprise BI and advanced analytics, now leads the pack in Agile BI.”

The **Forrester Wave™: Agile Business Intelligence Platforms, Q3 2014**

([link](#))
INTRODUCTION THE SAS SUPPORTED HADOOP DISTRIBUTIONS

- Cloudera
- Hortonworks
- Pivotal
- MapR
- IBM
INTRODUCTION

VISION FOR SAS ON HADOOP

To be the Analytic and Data Management solution of choice for Hadoop.
SAS AND HADOOP

TYPICAL MODERN ARCHITECTURE

SOURCE DATA
- Trade/Financial Feeds
- Sensor Data/Smart Device
- Telemetry
- Network Traffic
- Databases
- Routers, switches

SAS® REAL TIME DECISION MANAGER

LASR™ ANALYTICS SERVER
- Data In-Memory
- Extreme Parallelism
- Distribution of Analytics Processes

EVENT STREAM PROCESSING
- Querying
- Filtering
- Calculating
- Pattern Matching
- Aggregating

HADOOP CLUSTER

SAS® VISUAL STATISTICS

SAS® VISUAL ANALYTICS

SAS® IMSTAT FOR HADOOP
Even with In-Database processing there will still be some work performed on the SAS server.
HIGH LEVEL ARCHITECTURE

DISTRIBUTED DEPLOYMENT OF VA ON COMMODITY HARDWARE WITH ASYMMETRIC SOURCES

SAS® VISUAL ANALYTICS

DESKTOP CLIENTS
- SAS® Management Console

WEB-BASED CLIENTS
- Hub
- Explorer
- Designer
- Viewer
- Data Builder
- Administrator

MOBILE CLIENTS
- iPad
- Android

Hub
Explorer
Designer
Viewer
Data Builder
Administrator

WEB-BASED CLIENTS

Hadoop
RDBMS
Nonrelational
Click Stream
PC Files & more

1. Front loading through workspace server
2. Parallell load from hadoop (SASHDAT)
3. Parallell load from separate hadoop cluster using SAS EP

IN-MEMORY STORE
SAS® LASR ANALYTIC SERVER

WORKSPACE SERVER

MID-TO-TIER METADATA SERVER

Hadoop HDFS

Hadoop
RDBMS
Nonrelational
Click Stream
PC Files & more

SAS® LASR ANALYTIC SERVER CLUSTER

SAS LASR ANALYTIC SERVER

CLOUDERA / HORTONWORKS / TERADATA / GREENPLUM / DB2 / ORACLE / NETEZZA / SAP HANA

Not part of VA

Can be separated
**ARCHITECTURE OPTIONS**

**SHARED CLUSTER VS SEPARATE CLUSTERS**

**Shared Cluster**
- Resource Management (YARN)
- SASHDAT Available
- SAS Scales with Hadoop Cluster

**Separate clusters**
- Math scales separately from Hadoop
- Data “movement”
- No SASHDAT
WHAT IS THE SAS EMBEDDED PROCESS?

SAS EP

A portable, lightweight execution container for SAS code that makes SAS portable and deployable on a variety of platforms

1. Data Lifting
2. Data Preparation
3. Data Quality
4. Scoring
HOW DOES DATA LOADING IN ASYMMETRIC MODE WORK?

SAS® Visual Analytics

Data

Hadoop

EP
HADOOP

WHY IS YARN SO IMPORTANT?

 Map Reduce (MR)
  ➢ Job Tracker
    Submits tasks to the TaskTracker(s)
  ➢ Task Tracker
    Accepts Map, Reduce, and Shuffle Operations

 YARN
  ➢ a global ResourceManager (RM)
    Scheduling for cluster, queue capacities, user limits
  ➢ a per-application ApplicationMaster (AM)
    Negotiates resources from RM, works with NodeManager to execute and monitor component tasks
  ➢ a per-node slave NodeManager (NM) and
    a per-application Container running on a NM
    Launches application, monitors cpu, memory, disk, network and reports to RM

ARCHITECTURE HARDWARE BUILDING BLOCKS

- Node, Blade, Pizza Box, Rack Mount, Virtual Machine ….

- Components:
  1. RAM
  2. CPU
  3. Disks
  4. Network
  5. Linux x64

Sample:
20 cores (2 CPU)
512 GB RAM
16 x 2 TB SATA Disks
Dual 10GbE Ethernet
Red Hat Linux 6.5
ARCHITECTURE

ASYMERIC DEPLOYMENT

- Minimum configuration is four nodes in each cluster
- Recommended is six nodes for hadoop cluster
- No need for equal no of nodes
• Minimum configuration is four nodes
• Recommended is six
• In 9.4M3, yarn can manage the complete workload across hadoop, SAS Lasr and SAS Grid
• SAS can facilitate coordination with preferred Hadoop Vendor
• You need to buy separate license/support agreement from Hadoop vendor
• The vendors provide fixed priced installation of hadoop, not very expensive
• Needs to plan from the beginning due to differences in server configuration (disks)
**BASELINE SUPPORT MATRIX @ 9.4M2**

*Product Management makes every effort to validate that this is correct. However, the definitive answer to what is supported on what platforms, etc. can be found at INSTALL CENTER. Note the General forward support for SAS software.*

<table>
<thead>
<tr>
<th>9.4M2</th>
<th>Hive</th>
<th>Impala</th>
<th>Score Accl</th>
<th>Code Accl</th>
<th>DQ Accl</th>
<th>Data Ldr</th>
<th>SPDE</th>
<th>HPA</th>
<th>VA</th>
<th>IMSTAT</th>
<th>VS</th>
<th>VSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloudera 4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloudera 5.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HortonWorks 1.3.2</td>
<td></td>
<td></td>
<td>New 9.4M2</td>
<td>New 9.4M2</td>
<td>New 9.4M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HortonWorks 2.0</td>
<td></td>
<td></td>
<td>New 9.4M2</td>
<td>New 9.4M2</td>
<td>New 9.4M2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pivotal HD 1.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pivotal HD 2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM BigInsights 2.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MapR V3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Supported at 9.4M2**

**Not Supported**

* Minus Yarn Integration

**Net New Product**
### Support Matrix – as of 9.17.2014

Product Management makes every effort to validate that this is correct. However, the definitive answer to what is supported on what platforms, etc. can be found at INSTALL CENTER. Note the General forward support for SAS software.

<table>
<thead>
<tr>
<th></th>
<th>LASR Analytic Server 2.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HPA</td>
</tr>
<tr>
<td><strong>9.4M2</strong></td>
<td></td>
</tr>
<tr>
<td>Cloudera 4.5</td>
<td></td>
</tr>
<tr>
<td>Cloudera 5.0</td>
<td></td>
</tr>
<tr>
<td>HortonWorks 1.3.2</td>
<td></td>
</tr>
<tr>
<td>HortonWorks 2.0</td>
<td></td>
</tr>
<tr>
<td>Pivotal HD 1.1</td>
<td></td>
</tr>
<tr>
<td>Pivotal HD 2.0</td>
<td></td>
</tr>
<tr>
<td>IBM BigInsights 2.1</td>
<td>*</td>
</tr>
<tr>
<td>MapR V3</td>
<td></td>
</tr>
</tbody>
</table>

Supported at 9.4M2
Not Supported
* Minus Yarn Integration
Net New Product