QUEST meeting – Big Data Analytics

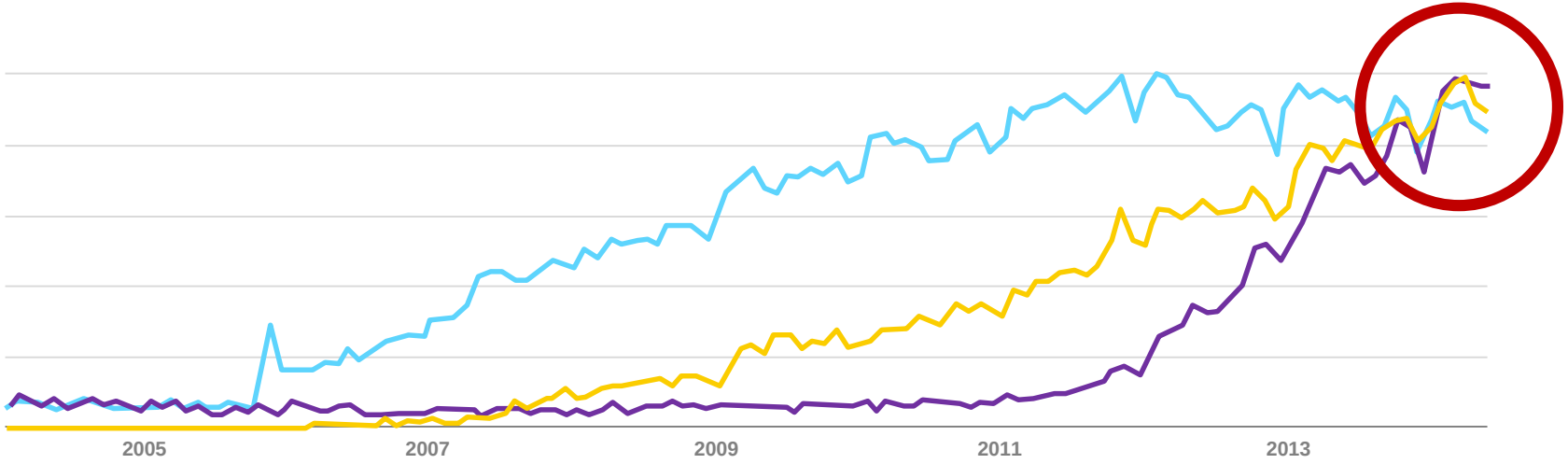
Peter Hughes
Business Solutions Consultant
SAS Australia/New Zealand

Copyright © 2015, SAS Institute Inc. All rights reserved.

Big Data Analytics

WHERE WE ARE NOW

Google trends



BIG DATA
Lots of data

HADOOP
Processing Power

ANALYTICS
Accurate /Decisions

The era of abundance

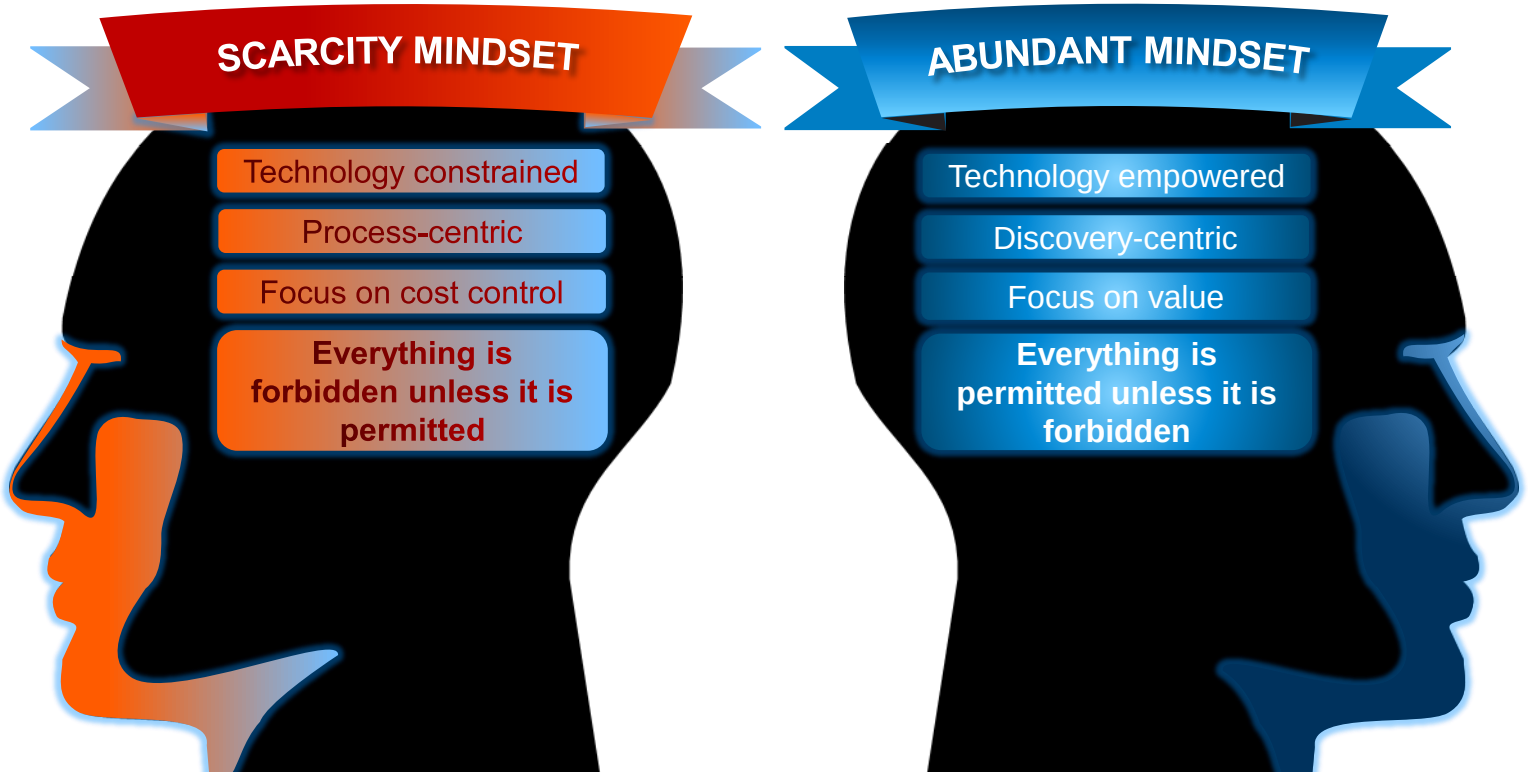
"Big data is what happened when the cost of storing information became less than the cost of making the decision to throw it away."

- George Dyson

Science Historian and TED Speaker

Copyright © 2014, SAS Institute Inc. All rights reserved.

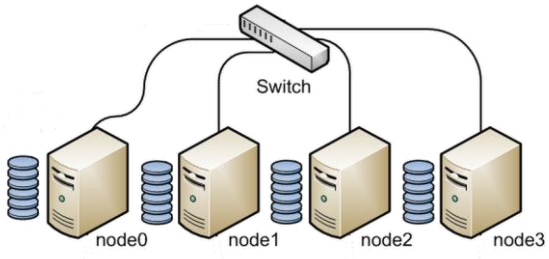
Two Eras . . . Will you modernize your mindset?



Copyright © 2014, SAS Institute Inc. All rights reserved.

WHAT IS HADOOP ?

- An Apache Software Foundation project
  - Open-source
  - Origins in early 2000s with contributions from Google, Yahoo! and Facebook
- Framework of tools for processing Big Data
  1. Base: Common, Distributed File System (HDFS); MapReduce & YARN
  2. Additional projects including: Pig; Hive; HBase; Pig; Zookeeper et al.
- Designed for clusters using commodity server hardware typically Intel/Linux
  - Distributed storage
  - Distributed processing
  - Fault-tolerant topology
- Commercial Hadoop distributions based on Apache code
  - Extensions; additional tooling; support
  - Vendors: Cloudera; Hortonworks, MapR; Pivotal; IBM; Intel & others



COMMERCIAL HADOOP VENDORS

Google Capital recently invested $80 Million to into MapR – they gathered $110 million of investment in their last round!

Intel recently invested $740 Million to buy 18%. Puts their value at around the $4 Billion mark!

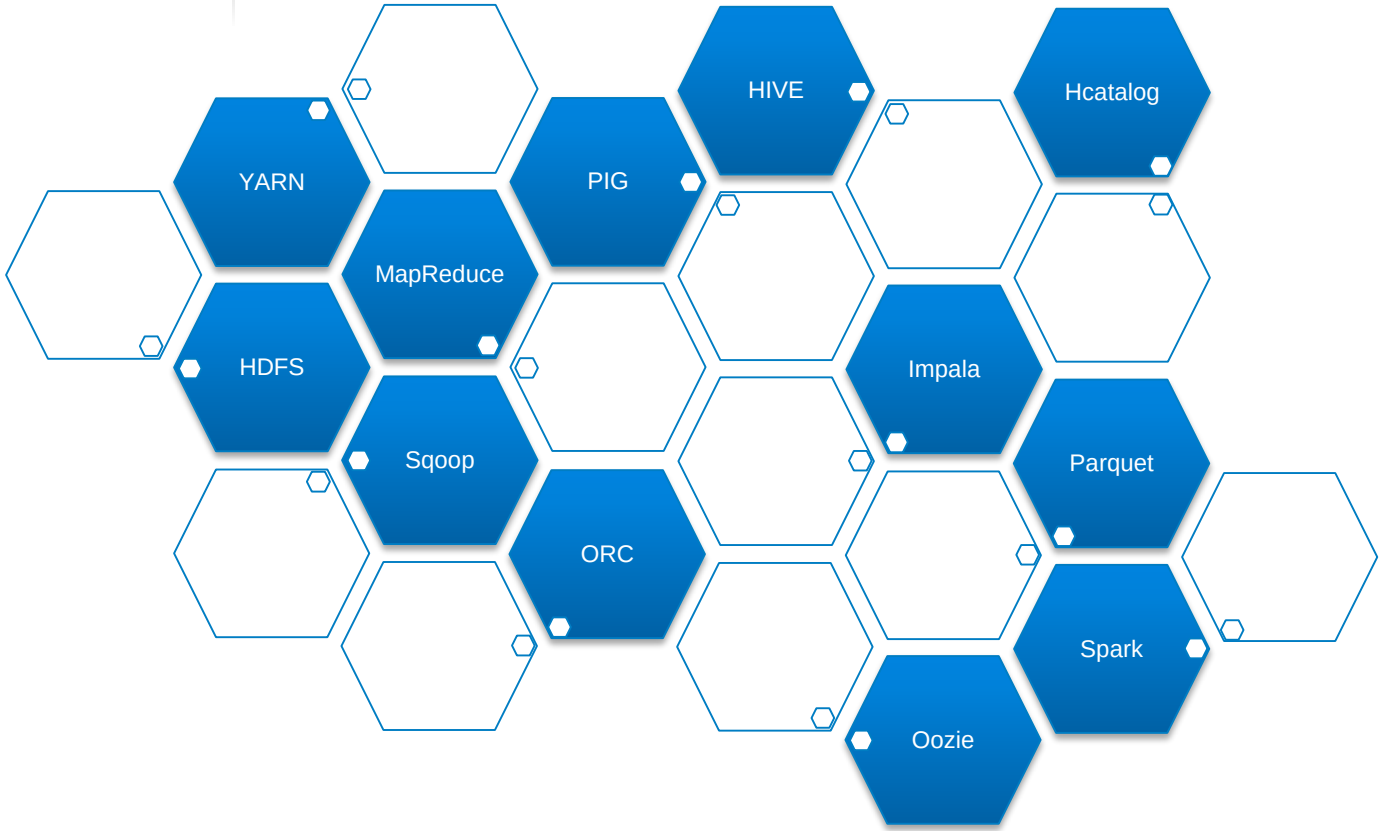
HP recently invested $50 Million to into Hortonworks to get a place on the board. Total investment now about $300 Million. Big Teradata and SAP Partners!

GE invested $105 Million In Pivotal

Pivotal HD

IBM InfoSphere BigInsights

INTEGRATION WITH OPEN SOURCE HADOOP

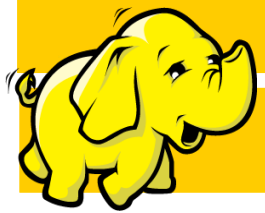


SAS® WITHIN THE HADOOP ECOSYSTEM

SAS® User

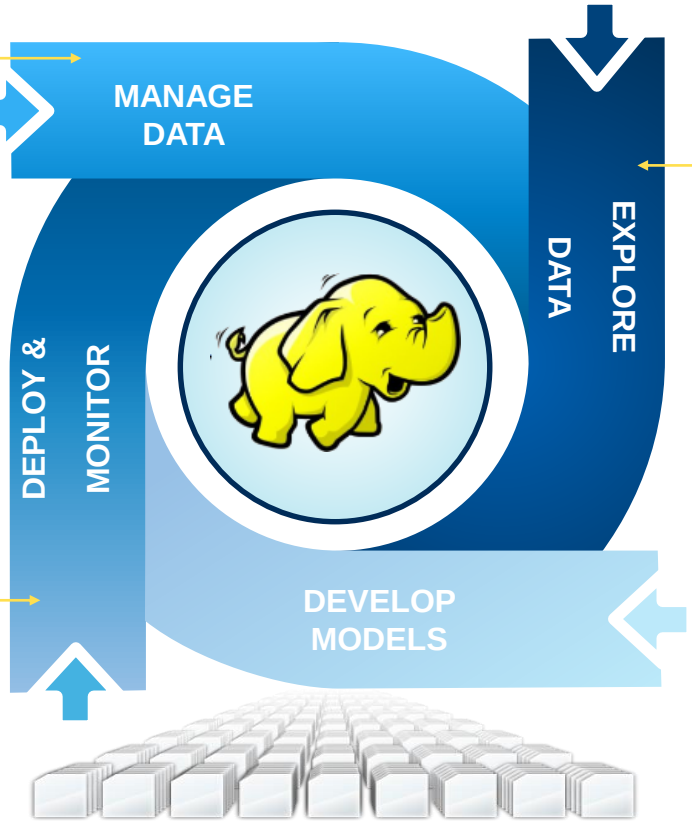
Next-Gen SAS® User

User Interface	SAS® Data Loader for Hadoop	SAS® Data Integration	SAS® Enterprise Miner™	SAS® Visual Analytics	SAS® In-Memory Statistics for Hadoop
Metadata	SAS Metadata				
Data Access	Base SAS & SAS/ACCESS® to Hadoop™			In-Memory Data Access	
Data Processing	Pig	Hive	SAS Embedded Process Accelerators	Server SAS® High-Performance Analytic Procedures	
	Map Reduce/YARN sas				
File System	HDFS				



DATA TO DECISION LIFECYCLE on Hadoop

- SAS/ACCESS (Hadoop/Impala)
  - SAS Data Management
  - SAS Federation Server
- SAS Data Quality Accelerator for Hadoop
- SAS Code Accelerator for Hadoop
- SAS Data Loader for Hadoop



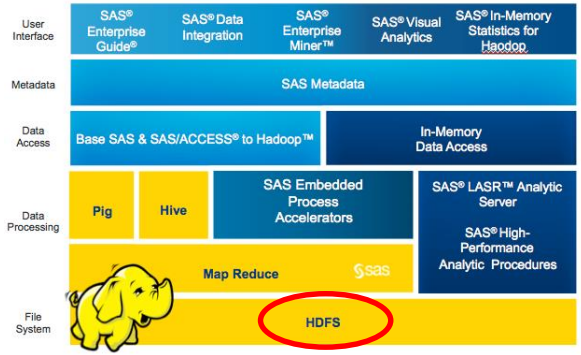
- SAS Visual Analytics
- SAS In-memory Statistics for Hadoop

- Model Manager
- SAS Scoring Accelerator for Hadoop

- SAS HPA Products
- *SAS Visual Statistics*
- SAS In-memory Statistics for Hadoop

Copyright © 2014, SAS Institute Inc. All rights reserved.

MANAGE DATA | READ/WRITE TO HDFS



file:///C:/Sample_data/hadoop_config.xml#

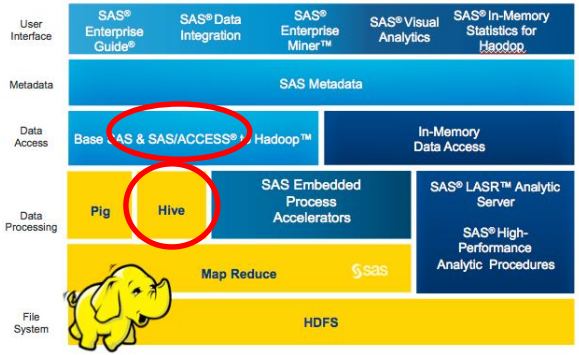
```
<?xml version="1.0" encoding="UTF-8"?>
- <configuration>
  - <property>
      <name>fs.default.name</name>
      <value>hdfs://sascldserv02:8020</value>
    </property>
  - <property>
      <name>mapred.job.tracker</name>
      <value>sascldserv02:8021</value>
    </property>
  - <property>
      <name>dfs.http.address</name>
      <value>sascldserv02:50070</value>
    </property>
  - <property>
      <name>dfs.secondary.http.address</name>
      <value>sascldserv02:50090</value>
    </property>
  - <property>
      <name>mapred.job.tracker.http.address</name>
      <value>sascldserv02:50030</value>
    </property>
</configuration>
```

```
/*
Create directory on HDFS
*/
filename cfg "C:\Sample_Data\hadoop_config.xml";
proc hadoop options=cfg username="hadoop" password="hadoop";
     hdfs mkdir="/user/hadoop/testfolder" ;
run;
/*
Copy file from local SAS to HDFS
*/
filename cfg "C:\Sample_Data\hadoop_config.xml";
proc hadoop options=cfg username="hadoop" password="hadoop";
     hdfs copyfromlocal="C:\Sample_data\dept.txt"
out="/user/hadoop/testfolder/";
run;
/*
Copy file from HDFS to local SAS
*/
filename cfg "C:\Sample_Data\hadoop_config.xml";
proc hadoop options=cfg username="hadoop" password="hadoop";
     hdfs copytolocal="/user/hadoop/testfolder" out="C:\Sample_data\" ;
run;
```

Hadoop configuration file, used for all `PROC HADOOP PIG|MAPREDUCE|HDFS` calls

Copyright © 2012, SAS Institute Inc. All rights reserved.

MANAGE DATA | SAS/ACCESS

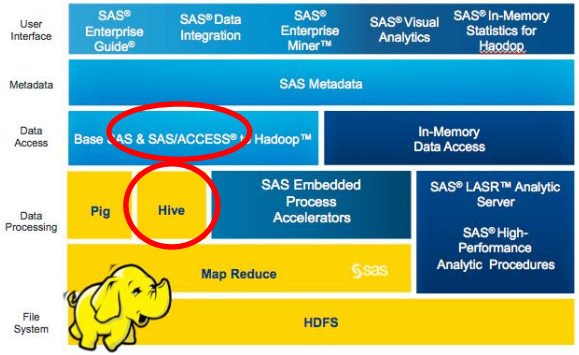


- Base SAS Procedures executed in-database for Hadoop
  - `FREQ, REPORT, SORT, SUMMARY/MEANS, TABULATE`
- Supported Hadoop distributions & combinations*
  - Cloudera CDH 5.0 running Hive/Hive2
  - Hortonworks HDP 2.0 running HiveServer2
  - IBM InfoSphere BigInsights 2.1 running Hive
  - MapR M5 2.0.1 running Hive
  - Pivotal/Greenplum HD running Hive
  - Pivotal/Greenplum MR 2.0.1 running Hive

* If a provider assures upward compatibility, SAS/ACCESS supports newer combinations. For example, Cloudera assures upward compatibility within major releases, so Cloudera CDH4.2 running Hive or HiveServer2 is supported.

Copyright © 2012, SAS Institute Inc. All rights reserved.

MANAGE DATA | HIVE

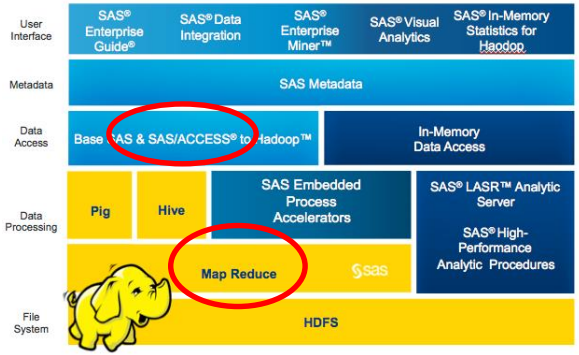



```
LIBNAME cdh_hdp HADOOP  PORT=10000 SERVER=sascldserv02 user=hadoop
password=hadoop ;
/*
Create new table
*/
proc sql;
connect to hadoop(PORT=10000 SERVER=sascldserv02  USER=hadoop
PASSWORD="hadoop");
exec( create table cars_prc (make string,  model string,  msrp double) ) by
hadoop;
quit;

/*
Copy from another table
*/
proc sql;
  insert into cdh_hdp.cars_prc
  select make, model, msrp
  from sashelp.cars ;
quit;
/*
List contents
*/
proc sql;
select * from cdh_hdp.cars_prc;
quit;
```

Copyright © 2012, SAS Institute Inc. All rights reserved.

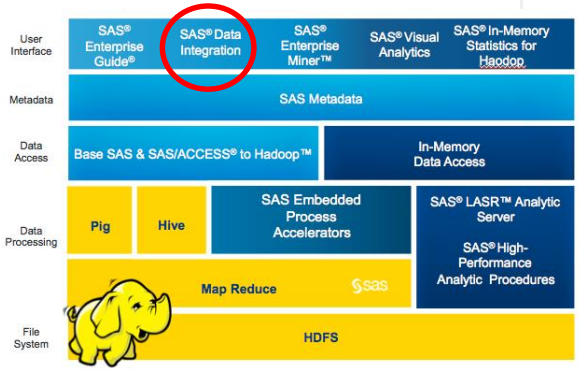
MANAGE DATA | MAPREDUCE



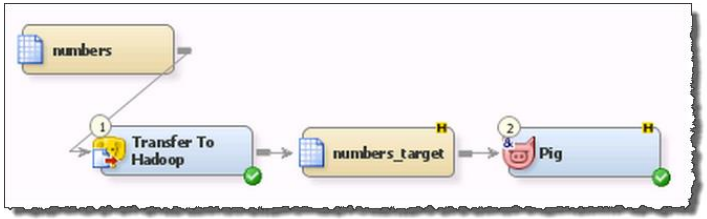

```
/*
Invoke MapReduce Word Count program
*/
filename cfg "C:\Sample_Data\hadoop_config.xml";
proc hadoop options=cfg username="hadoop" password="hadoop" verbose;
hdfs delete="/user/hadoop/output_MR1";
     mapreduce
     input="/user/hadoop/gutenberg" output="/user/hadoop/output_MR1"
     jar="C:\Sample_data\hadoop-examples-2.0.0-mr1-cdh4.1.2.jar"
     outputkey="org.apache.hadoop.io.Text"
     outputvalue="org.apache.hadoop.io.IntWritable"
     reduce="org.apache.hadoop.examples.WordCount$IntSumReducer"
     combine="org.apache.hadoop.examples.WordCount$IntSumReducer"
     map="org.apache.hadoop.examples.WordCount$TokenizerMapper"
     reducetasks=0
      ;
run;
```

Copyright © 2012, SAS Institute Inc. All rights reserved.

MANAGE DATA | SAS DATA INTEGRATION STUDIO



- Seamless access to Hadoop data (HDFS/HIVE/IMPALA) by analyst/traditional SAS users
- Reading & writing to/from HDFS
- Transfer to/from Hadoop operators
- Support for Pig, Hive & MapReduce transforms

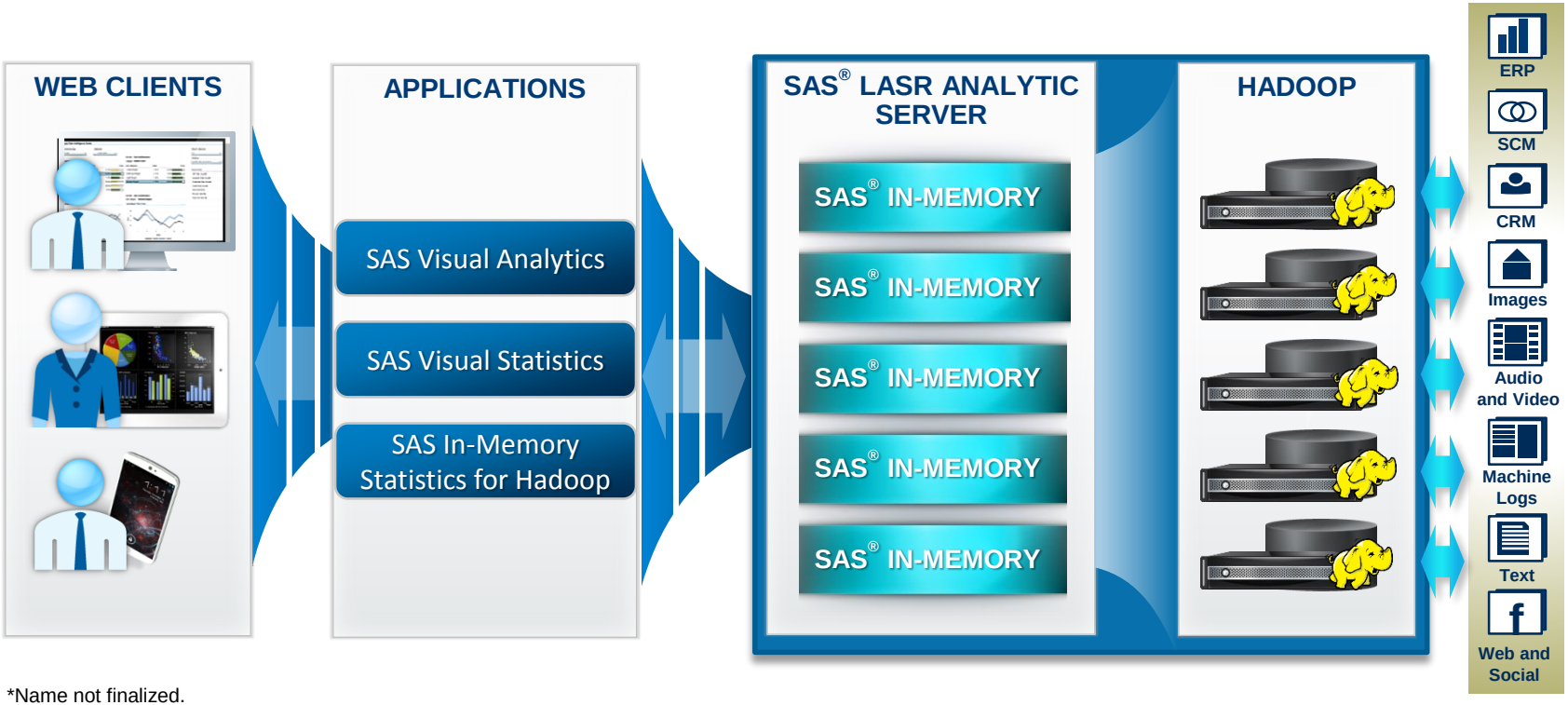


Copyright © 2012, SAS Institute Inc. All rights reserved.

SAS® IN-MEMORY ANALYTICS

SAS® LASR ANALYTIC SERVER AND HADOOP

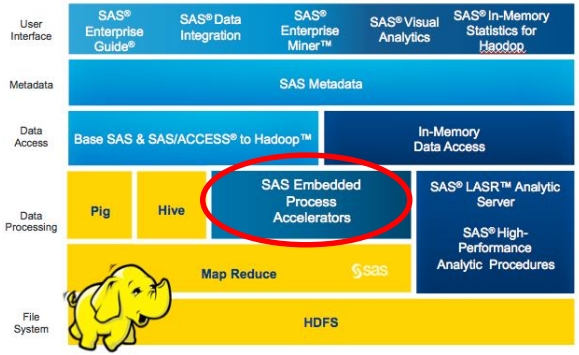
In-memory processing; use Hadoop for storage persistence and commodity computing



*Name not finalized.

Copyright © 2012, SAS Institute Inc. All rights reserved.

DEPLOY & MONITOR | SAS SCORING ACCELERATOR FOR HADOOP



- Publish SAS® Enterprise Miner™ models or SAS/STAT linear models inside the Hadoop
- Fully integrated with SAS® Model Manager to streamline registration, validation and performance monitoring
- Reduced data movement and improve data governance by streamlining model deployment processes within Hadoop

Copyright © 2012, SAS Institute Inc. All rights reserved.

http://www.sas.com/au/sashadoop

Copyright © 2014, SAS Institute Inc. All rights reserved.

QUESTIONS?

Copyright © 2012, SAS Institute Inc. All rights reserved.

Peter.hughes@sas.com

peter hughes

Thank You!

http://www.sas.com/au/sashadoop

Copyright © 2012, SAS Institute Inc. All rights reserved.