HADOOP FOR SAS ADMINISTRATORS

DOUG GREEN, SAS UK

HADOOP FOR SAS ADMINISTRATORS

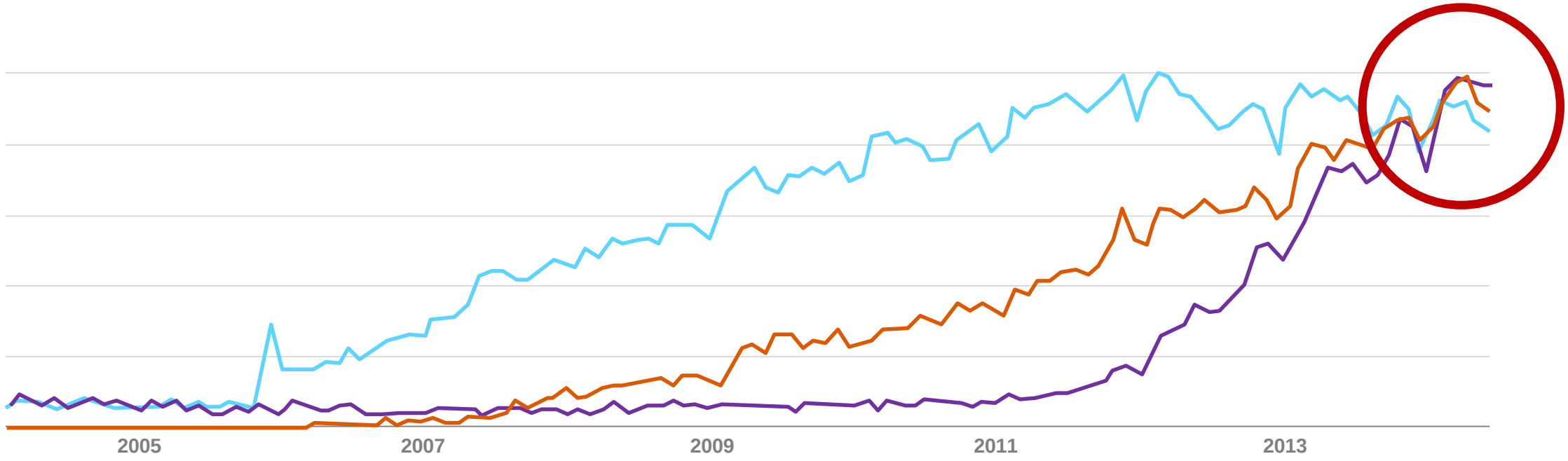
AGENDA

- Why your companies will move to Hadoop
- Why you should consider Hadoop and SAS
- Practical elements

HADOOP FOR SAS ADMINISTRATORS

WHERE WE ARE NOW

Google trends



BIG DATA

Lots of data

HADOOP

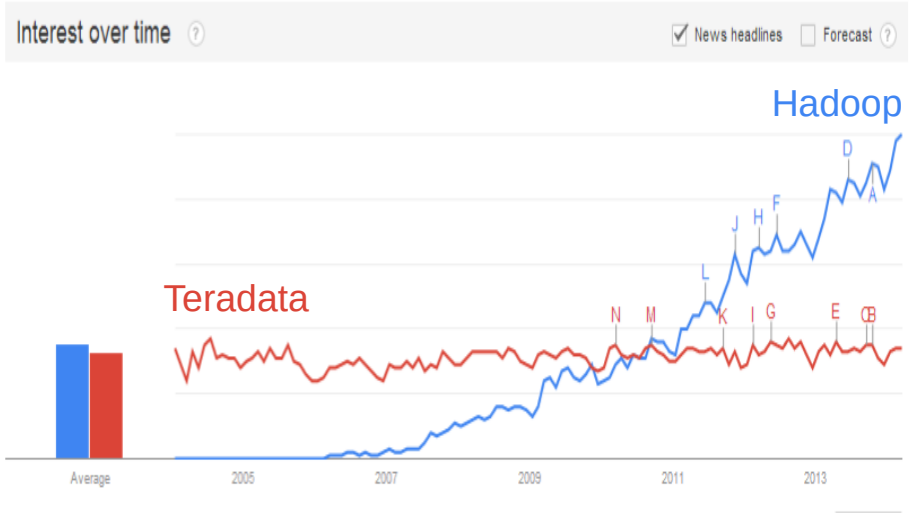
Processing Power

ANALYTICS

Accurate Decisions

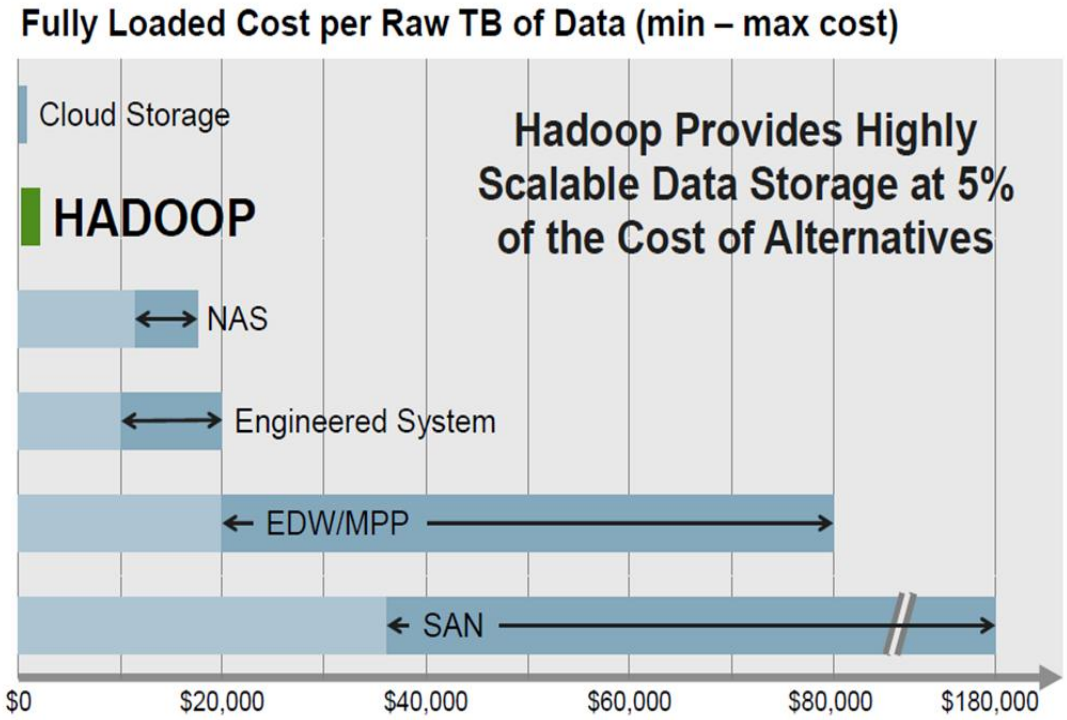
HADOOP FOR SAS ADMINISTRATORS

WHY SHOULD YOU CARE?



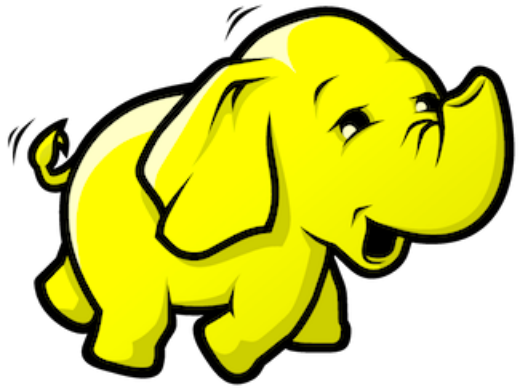
WHY YOUR COMPANY WILL MOVE TOWARDS HADOOP

- Lower Cost
  - Cost of storage
  - Cost of processing
  - Store data without a schema (schema on read)
- Greater Opportunity
  - Manage & analyse unstructured & structured data
  - Advanced analytics at scale



HADOOP – THE ~~FUTURE~~ BIG DATA PLATFORM

Challenges
Ecosystem maturity Skills shortage Tools / Application Shortage



Benefits
Cost Agility Speed

SAS View:
Simplify the skills and tooling issue. Ecosystem maturing quickly

SAS View:
More data at a lower cost.
Innovate like a start-up.
Speed as an enabler

HADOOP FOR SAS ADMINISTRATORS

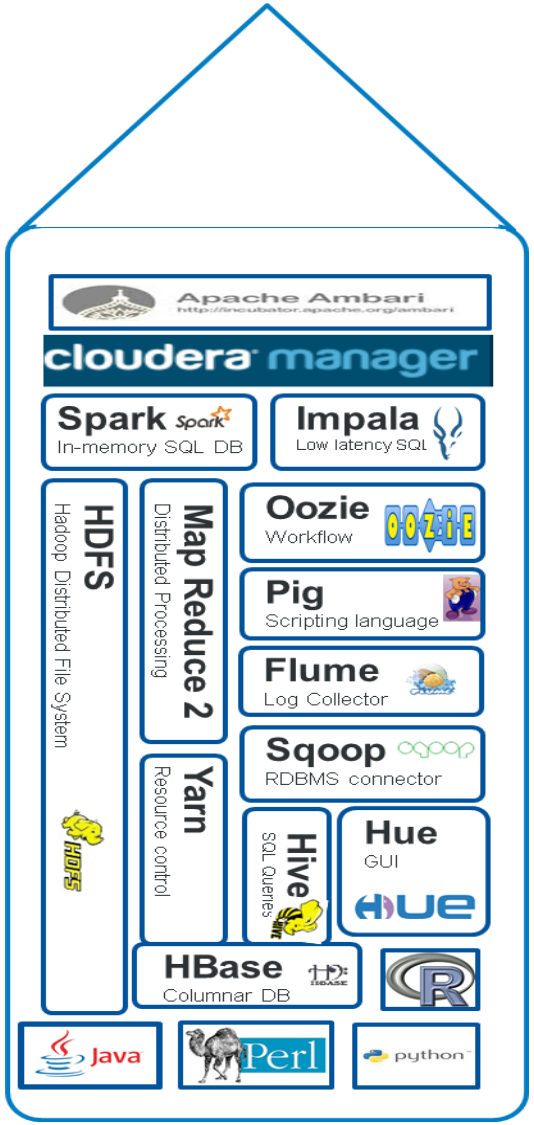
HADOOP ECOSYSTEM AND REQUIRED SKILLS

HIVEQL

```
SELECT `t1`.`i`,
       `t1`.`age`,
       `t2`.`salary`,
       `t3`.`ltv`
      FROM `TABLE1` `t1`, `TABLE2` `t2`, `TABLE3` `t3`
      WHERE (`t1`.`i` = `t2`.`i` AND `t1`.`i` = `t3`.`i`) AND
`t2`.`salary` < 20000);
```

MapReduce

```
public class WordCount {

  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new
IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context
context
                    ) throws IOException,
InterruptedException {
      StringTokenizer itr = new
StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
```



PIG

```
"RAW = LOAD '/user/sukdmg/RawFiles/Hamlet' as
(L:chararray); ";
"F = FILTER RAW BY (($0) MATCHES '""[A-Z]*'); ";
"GG = FOREACH F GENERATE L ; ";
"GGT = FOREACH GG GENERATE TOKENIZE(L) ; ";
"GGTL = FOREACH GGT GENERATE FLATTEN ($0) ; ";
"GGTG = GROUP GGTL BY ($0) ; ";
"GGTC = FOREACH GGTG GENERATE group, COUNT(GGTL)
; ";
"STORE GGTC INTO
'/demo/generic/raw_data/HamletNames.txt' USING
PigStorage('\t');";
```

SPARK (via SCALA)

```
 val points =
spark.textFile(...).map(parsePoint).cache()
var w = Vector.random(D) // current separating
plane
for (i <- 1 to ITERATIONS) {
   val gradient = points.map(p =>
     (1 / (1 + exp(-p.y*(w dot p.x))) - 1) *
p.y * p.x
   ).reduce(_ + _)
   w -= gradient
 }
 println("Final separating plane: " + w)
```

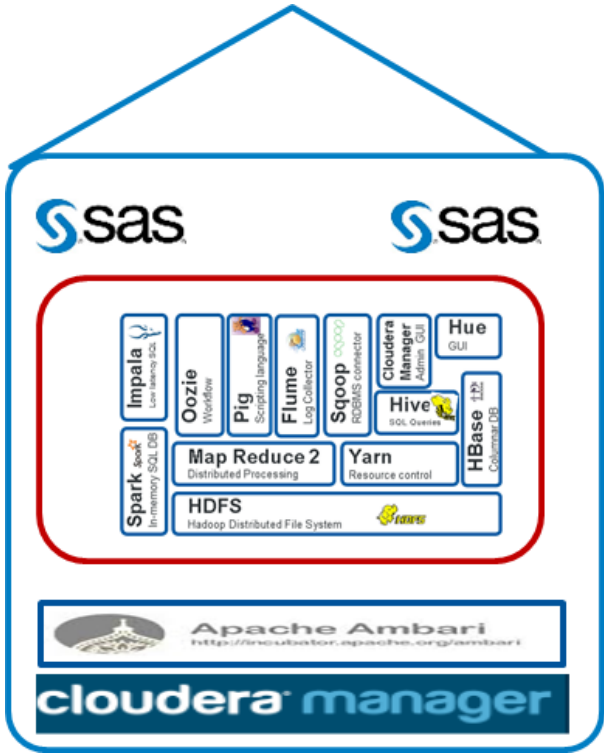
HADOOP FOR SAS ADMINISTRATORS

SKILLS REQUIRED WITH SAS & HADOOP

SAS Data Integration Studio 4.9

SAS® VISUAL ANALYTICS

DS1, DS2



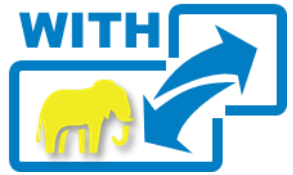
Implicit & explicit

HOW SAS & HADOOP INTEGRATE

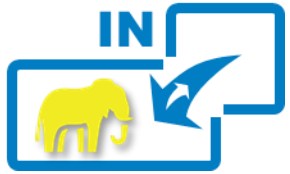
SAS & Hadoop intersect in many ways:

- SAS can treat Hadoop just as any other data source, pulling data **FROM** Hadoop, when it is most convenient;

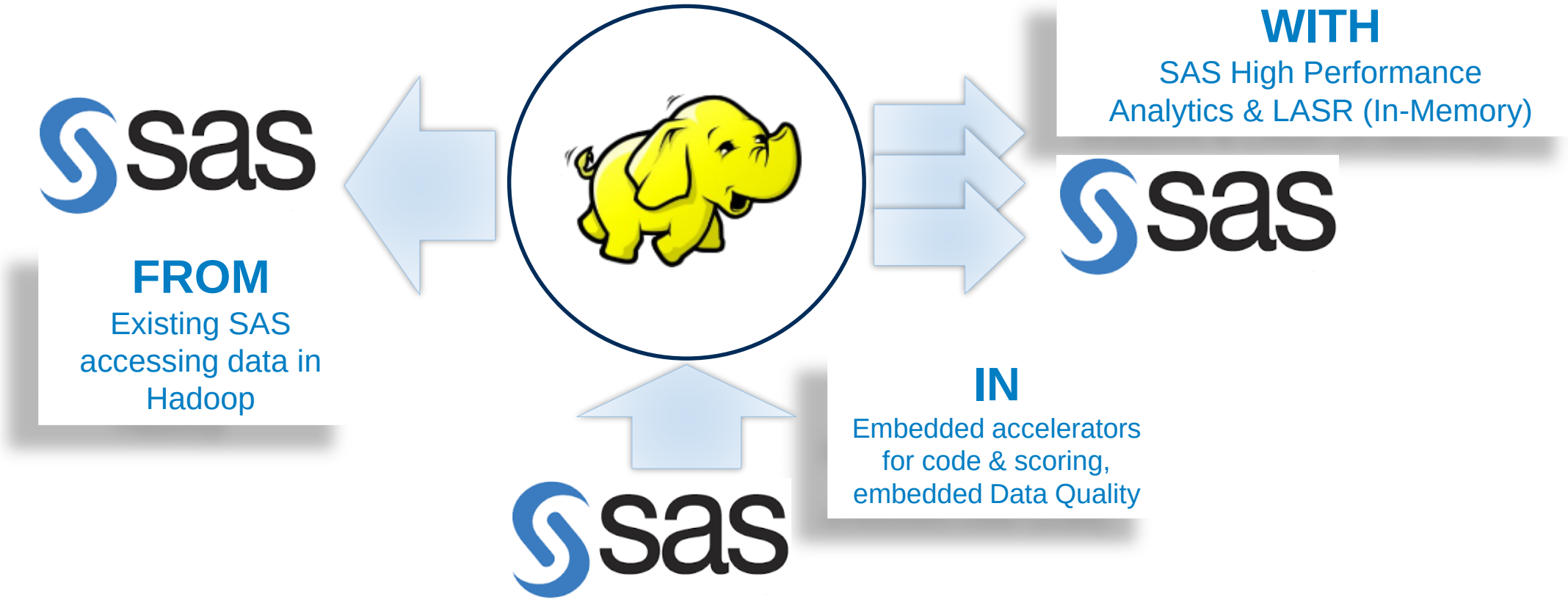
- SAS can work **WITH** Hadoop, lifting data into a purpose-built advanced analytics in-memory environment;



- SAS can work directly **IN** Hadoop, leveraging the distributed processing capabilities of Hadoop.



HOW SAS & HADOOP INTEGRATE



PRACTICAL ELEMENTS

- Useful skills for working with SAS and Hadoop
- How to set up SAS/Access to Hadoop (brief overview)
- “It doesn’t work…” – typical gotchas from users
- What to try next

USEFUL SKILLS FOR SAS & HADOOP

- SQL (but not necessarily ANSI standard SQL….)
- General SAS programming knowledge, DS1, DS2, SAS Procs etc.
- SAS/Access concepts (implicit & explicit pass-through)

- Linux skills (Linux on x86 is the de facto OS for Hadoop)
- HDFS security (think Linux: Owner, Group, World)
- Hadoop command line tools (Beeline, Grunt etc.)
- Hue (Web UI for Hadoop)
- Hadoop Admin interfaces (Cloudera Manager, Ambari)
- YARN awareness

HOW TO SET UP SAS/ACCESS TO HADOOP (OVERVIEW)

1. Copy Hadoop Jar files to SAS environment
2. Copy Hadoop XML files to SAS environment
3. Set environment variables
    1. SAS_HADOOP_JAR_PATH
    2. SAS_HADOOP_CONFIG_PATH
4. Test the libnames
5. Register Hadoop Cluster in Metadata
6. Create HIVE libnames in Metadata
7. Register HIVE data in Metadata

COPY JAR FILES

/data/SAS/config/HadoopConfig/JARS

Name	Ext	Size	Changed	Rights	Owner
..			27/02/2015 14:51:39	rwxr-xr-x	sasinst
MR1			25/02/2015 19:30:04	rwxr-xr-x	sasinst
activation-1.1.jar		62,983 B	25/02/2015 19:30:03	rw-r--r--	sasinst
apacheds-i18n-2.0.0-M15.jar		44,925 B	25/02/2015 19:30:03	rw-r--r--	sasinst
apacheds-kerberos-codec-2.0.0-M15.jar		675 KiB	25/02/2015 19:30:03	rw-r--r--	sasinst
api-asn1-api-1.0.0-M20.jar		16,560 B	25/02/2015 19:30:03	rw-r--r--	sasinst
api-util-1.0.0-M20.jar		79,912 B	25/02/2015 19:30:03	rw-r--r--	sasinst
asm-3.2.jar		43,398 B	25/02/2015 19:30:03	rw-r--r--	sasinst
avro-1.7.6-cdh5.3.1.jar		426 KiB	25/02/2015 19:30:03	rw-r--r--	sasinst
aws-java-sdk-1.7.4.jar		11,668 KiB	25/02/2015 19:30:03	rw-r--r--	sasinst
commons-beanutils-1.7.0.jar		184 KiB	25/02/2015 19:30:03	rw-r--r--	sasinst

http://support.sas.com/resources/thirdpartysupport/v94/hadoop/hadoopbacg.pdf

COPY XML CONFIGURATION FILES

/data/SAS/config/HadoopConfig/XML

Name	Ext	Size	Changed	Rights	Owner
..			27/02/2015 14:51:39	rwxr-xr-x	sasinst
merged			02/03/2015 07:41:40	rwxr-xr-x	sasinst
CDH511.tar.gz		2,318 B	14/01/2015 12:29:59	rw-r--r--	root
XML64M2.tar.gz		2,337 B	23/12/2014 15:27:47	rw-r--r--	root
config.xml		15,119 B	14/01/2015 12:42:22	rw-r--r--	sasinst
core-site.xml		3,368 B	03/02/2015 16:55:14	rw-r--r--	sasinst
hdfs-site.xml		1,654 B	03/02/2015 16:55:14	rw-r--r--	sasinst
hive-site.xml		2,195 B	03/02/2015 16:55:50	rw-r--r--	sasinst
mapred-site.xml		4,593 B	24/02/2015 16:44:58	rw-r--r--	sasinst
yarn-site.xml		3,941 B	03/02/2015 16:55:50	rw-r--r--	sasinst

http://support.sas.com/resources/thirdpartysupport/v94/hadoop/hadoopbacg.pdf

HADOOP FOR SAS ADMINISTRATORS

SET ENVIRONMENT VARIABLES

/data/SAS/config/Lev1/SASApp

Name	Ext	Size	Changed	Rights	Owner
..			05/03/2015 01:06:12	rwxr-xr-x	sasinst
BatchServer			24/08/2014 17:44:56	rwxr-xr-x	sasinst
ConnectServer			05/12/2014 08:39:06	rwxr-xr-x	sasinst
Data			24/08/2014 17:52:34	rwxr-xr-x	sasinst
PooledWorkspaceServer			13/01/2015 08:16:57	rwxr-x---	sasinst
SASEnvironment			24/08/2014 17:43:15	rwxr-xr-x	sasinst
StoredProcessServer			24/08/2014 17:45:39	rwxr-x---	sasinst
WorkspaceServer			24/08/2014 17:46:03	rwxr-xr-x	sasinst
sasv9.cfg		1,171 B	24/08/2014 17:43:13	rw-r--r--	sasinst
sasv9_usermods.cfg		295 B	24/08/2014 17:43:14	rw-r--r--	sasinst
appserver_autoexec.sas		338 B	24/08/2014 17:43:13	rw-r--r--	sasinst
appserver_autoexec_usermods.sas		436 B	09/09/2014 16:12:43	rw-r--r--	sasinst
appservercontext_env.sh		542 B	12/01/2015 14:15:25	rw-r--r--	sasinst
appservercontext_env_usermods.sh		203 B	24/08/2014 17:43:14	rwxr-xr-x	sasinst
sas.sh		894 B	24/08/2014 17:43:13	rwxr-xr-x	sasinst
sas_usermods.sh		165 B	24/08/2014 17:43:14	rwxr-xr-x	sasinst

/data/SAS/config/Lev1/SASApp/appserver_autoexec_usermods.sas - HPA2 head node - Editor - WinSCP

Encoding

```
/*
 * appserver_autoexec_usermods.sas
 *
 *    This autoexec file extends appserver_autoexec.sas.  Place your site-specific include
 *    statements in this file.
 *
 *    Do NOT modify the appserver_autoexec.sas file.
 *
 */

/* Following added by Pete on request of Doug 09SEP14 */

option set=SAS_HADOOP_CONFIG_PATH="/data/SAS/config/HadoopConfig/XML";
option set=SAS_HADOOP_JAR_PATH="/data/SAS/config/HadoopConfig/JARS";
```

Line: 1/15 | Column: 1 | Character: 47 (0x2F) | Encoding: 1252 (ANSI - Latin

HADOOP FOR SAS ADMINISTRATORS

TEST HADOOP LIBNAMES THROUGH CODE

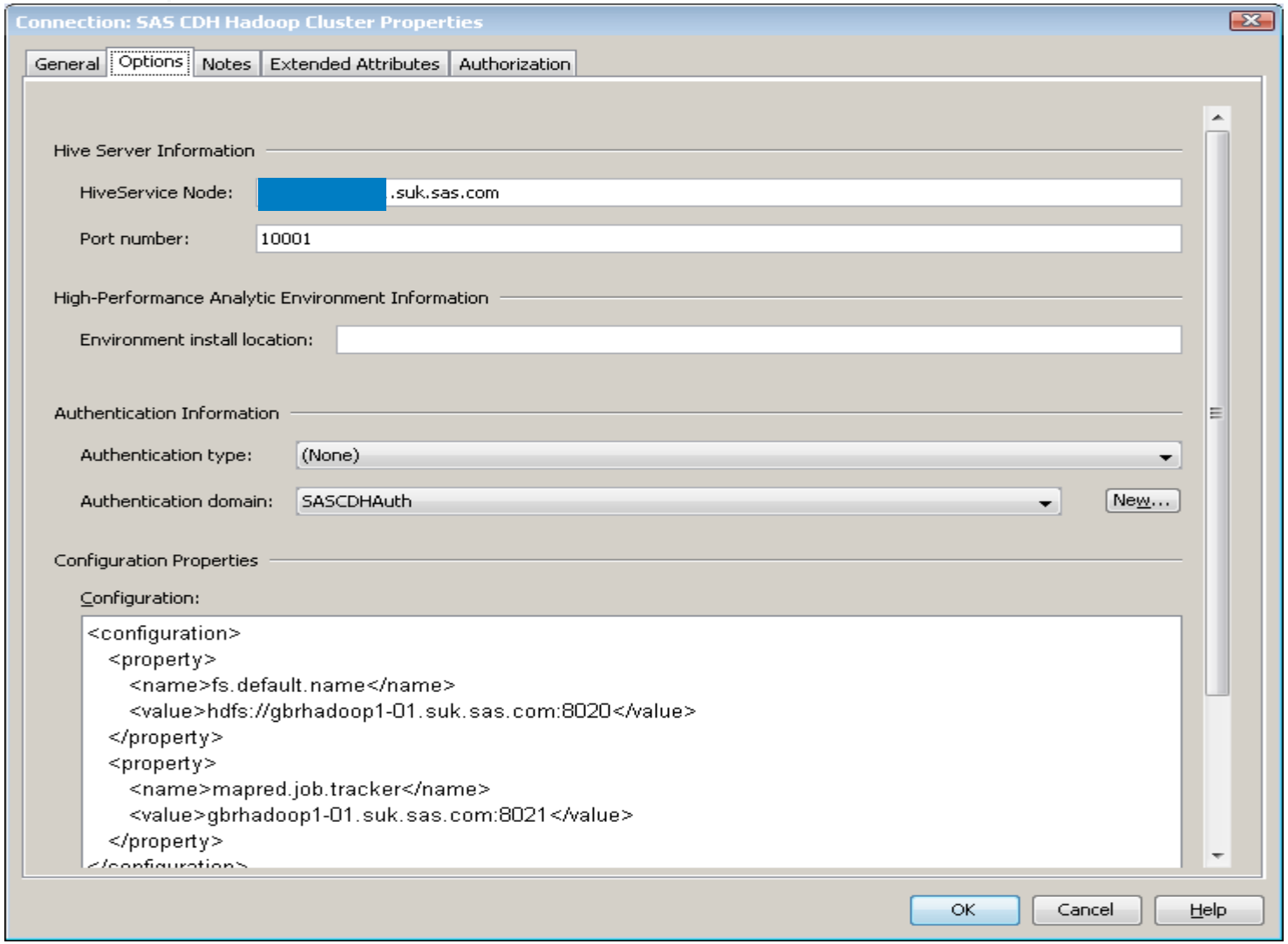
Program* | Log

Save | Run | Stop | Selected Server: SASApp (Connected) | Analyze Program | Export | Send To | Create | Properties

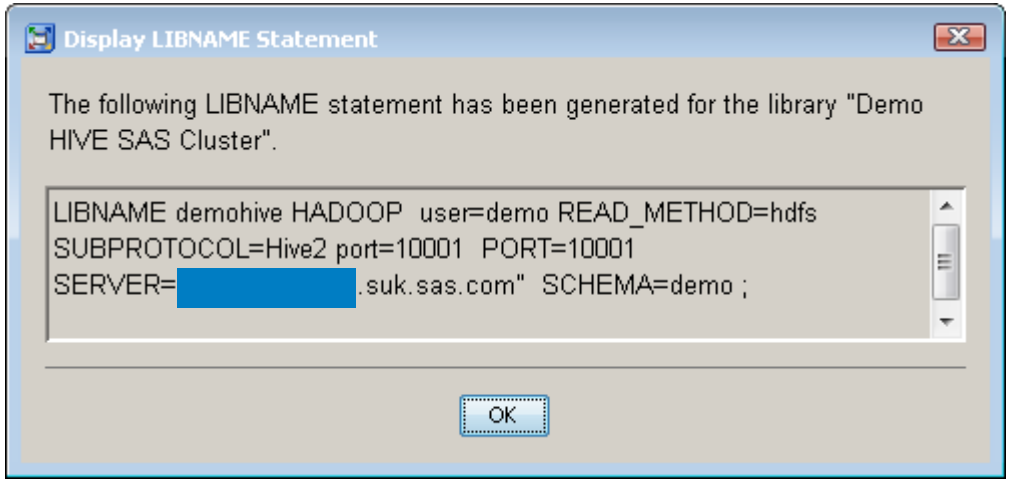
```
options symbolgen;

%let HADOOP_NAME_NODE=XXXXXXXX-01.suk.sas.com; *<--------the hadoop name node;
%let HIVE_SCHEMA=sukdmg; *<--------the HIVE schema you want to connect too;
%let HIVE_PORT=10001; *<--------the port number that the hive service is running on (default is 10000);
%let USER=sukdmg; *<--------the userid to connect with;
%let PWD={SAS002}E043FE4757B4CE074DC2458F2E9204C53282784D2A0DA252; *<--------the password to connect with;

%let IMPALA_HOST=XXXXXXXXX-02.suk.sas.com; *<--------the hadoop host name running the Impala gateway;
%let IMPALA_PORT=21050; *<--------the hport number that imapala is listening on (default is 21050;

%let HDFS_TMP=/user/sukdmg/anyfile/tmp;  *<--------hdfs TMP area;
%let HDFS_META=/user/sukdmg/anyfile/metadata; *<--------location where you want to store HDFS metadata files;
%let HDFS_RAW=/user/sukdmg/anyfile/raw;  *<--------location of raw HDFS files;

%let SPDE_HDFS_PATH=/user/sukdmg/spde; *<--------HDFS path for SPDE files;


/*HIVE libname */
libname sashive hadoop SUBPROTOCOL=hive2 READ_METHOD=HDFS schema=&HIVE_SCHEMA user=&USER pwd="&PWD"
server="&HADOOP_NAME_NODE" port=&HIVE_PORT ;

/*HDFS libname */
libname sashdfs hadoop schema=&HIVE_SCHEMA user=&USER pwd="&PWD" server="&HADOOP_NAME_NODE"
hdfs_tempdir = "&HDFS_TMP"
hdfs_metadir = "&HDFS_META"
hdfs_permdir = "&HDFS_RAW" ;

/*Impala Libname - Cloudera only */
LIBNAME sasimp SASIOIMP server="&IMPALA_HOST" schema=&HIVE_SCHEMA port=&IMPALA_PORT;

/*SPDE on HDFS libname example*/
libname spdehdfs spde "&SPDE_HDFS_PATH" hdfshost=default ACCELWHERE=YES;
```

HADOOP FOR SAS ADMINISTRATORS

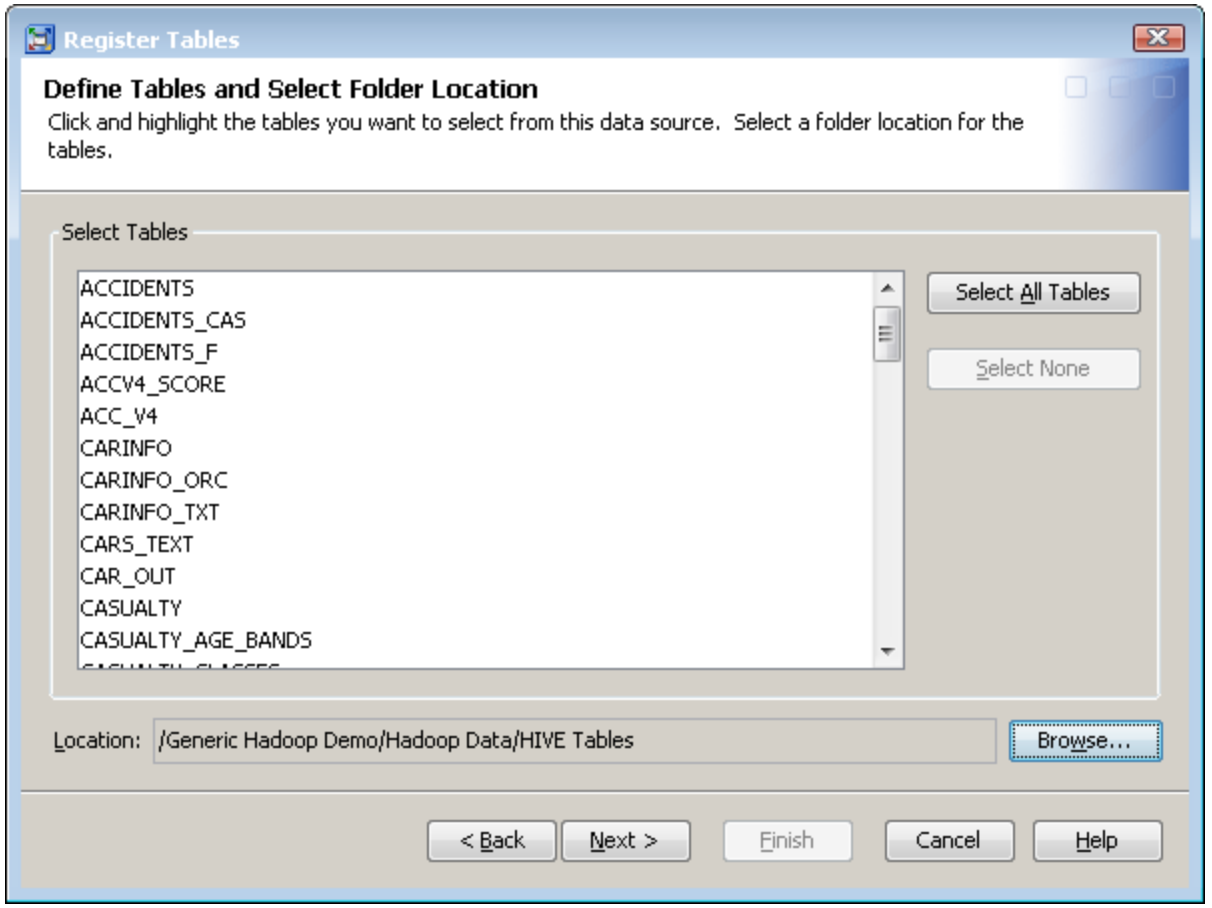
CREATE HADOOP CLUSTER IN METADATA



CREATE HIVE LIBNAME IN METADATA



REGISTER HIVE TABLES



TYPICAL HADOOP GOTCHAS FROM SAS USERS

- HIVE can’t handle implicit joins (can in HIVE 0.13 +)
  - Need to edit code generated by SAS Enterprise Guide query builder
  - Can specify the join in Data Integration (e.g. Create Table transformation)
- Testing “push-down” in DI requires the “explain” privilege in Hive

TYPICAL GOTCHAS FROM USERS

- HIVE Data Conversion issues
  - HIVE String types have a default format of SAS character $32767.
  - Always check you code with 10 or 20 obs into a dummy (use obs=10 dataset option) set to see if any variables are being read in at 32k
  - Can add additional HIVE metadata so that SAS reads correctly:

```
proc sql;
connect to hadoop (user="xxx" pw="*********" server=yyy port=10000 schema=default);
Execute (ALTER TABLE xxxx set TBLPROPERTIES ('SASFMT:account_source_cd'='char(100)',
'SASFMT:policy_quote_ref'='char(100)') )
BY HADOOP;
```

  - Another useful dataset or libname option is dbmax_text=5 which truncates all character variables to length 5
  - In HIVE 0.12+ the VARCHAR data type supported – always use this instead of STRING

WHY YOU SHOULD CONSIDER HADOOP AND SAS

- Get ahead of the game if Hadoop is not already part of your IT strategy…
  - It's probably coming anyway!
  - Work out what Hadoop is good and bad at for your organisation
  - Identify potential use cases and cost savings

NEXT STEPS

- Download Hadoop sandbox VM's and have a play
  - Hortonworks
  - Cloudera
- Migrate to SAS 9.4!!!
  - Trial the SAS Access engine(s)
  - Learn DS2
- Try the 90 day free trial of the Data Loader for Hadoop (out now!)

http://www.sas.com/en_us/software/data-management/data-loader-hadoop.html

SAS RESOURCES

- http://www.sas.com/en_us/insights/big-data/hadoop.html

HADOOP RESOURCES

- http://cloudera.com/content/cloudera/en/training/library/hadoop-essentials.html
- http://hortonworks.com/get-started/

QUESTIONS?