# SAS Platform Admin User group

*A glimpse of what we've learnt when we migrated to SAS 9.4 with connectivity to Hadoop*

*Date : 23 Feb and 2nd March 2016*

*Author: Malcolm Thorn / John Heathcote*

British Gas
Looking after your world

© British Gas

# Topics

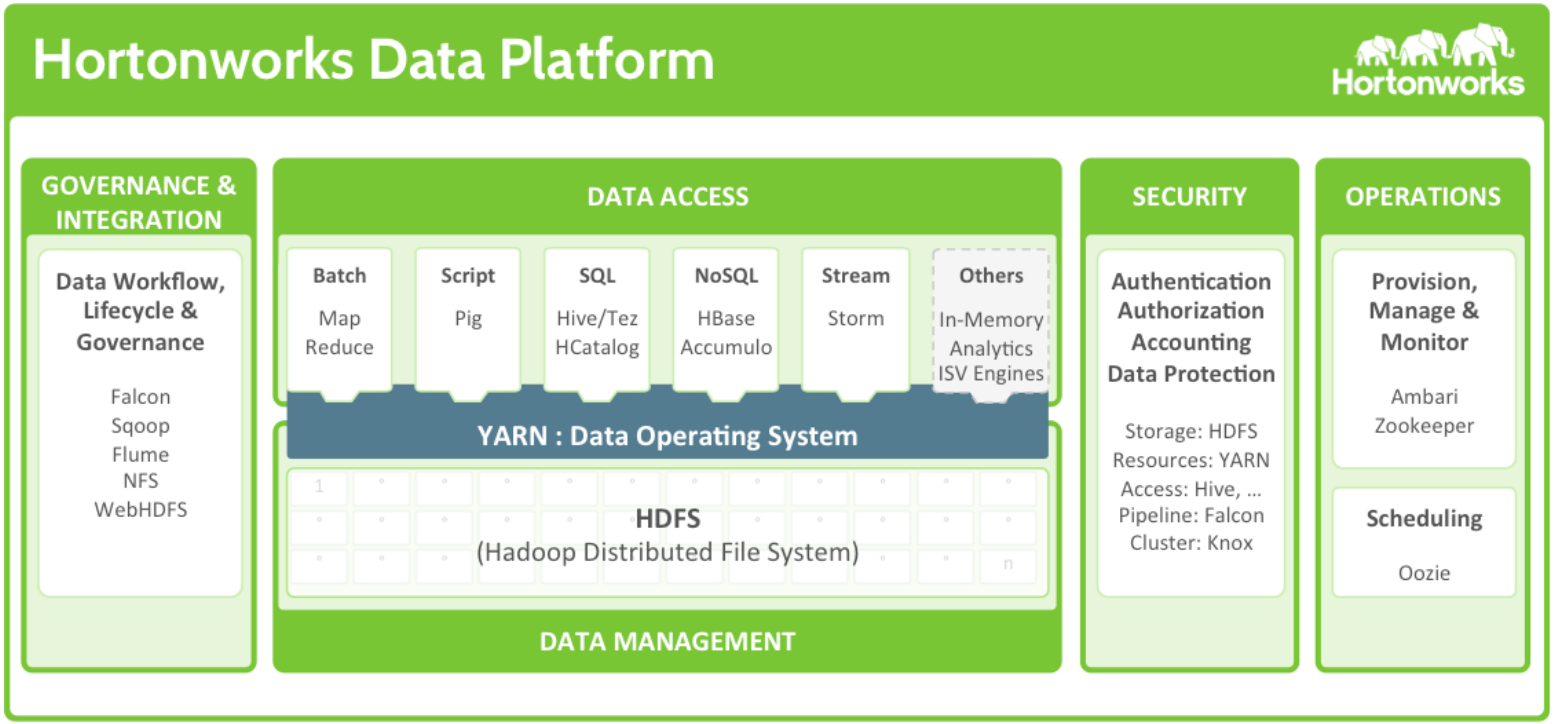
- What is Hadoop, why Hadoop?
- Migration - setup approach & challenges
- Top Admin Issues
- Top User Issues
- Benefits / success stories
- Top 5 recommendations and Summary
- Questions

© British Gas

# What is Hadoop, why Hadoop?

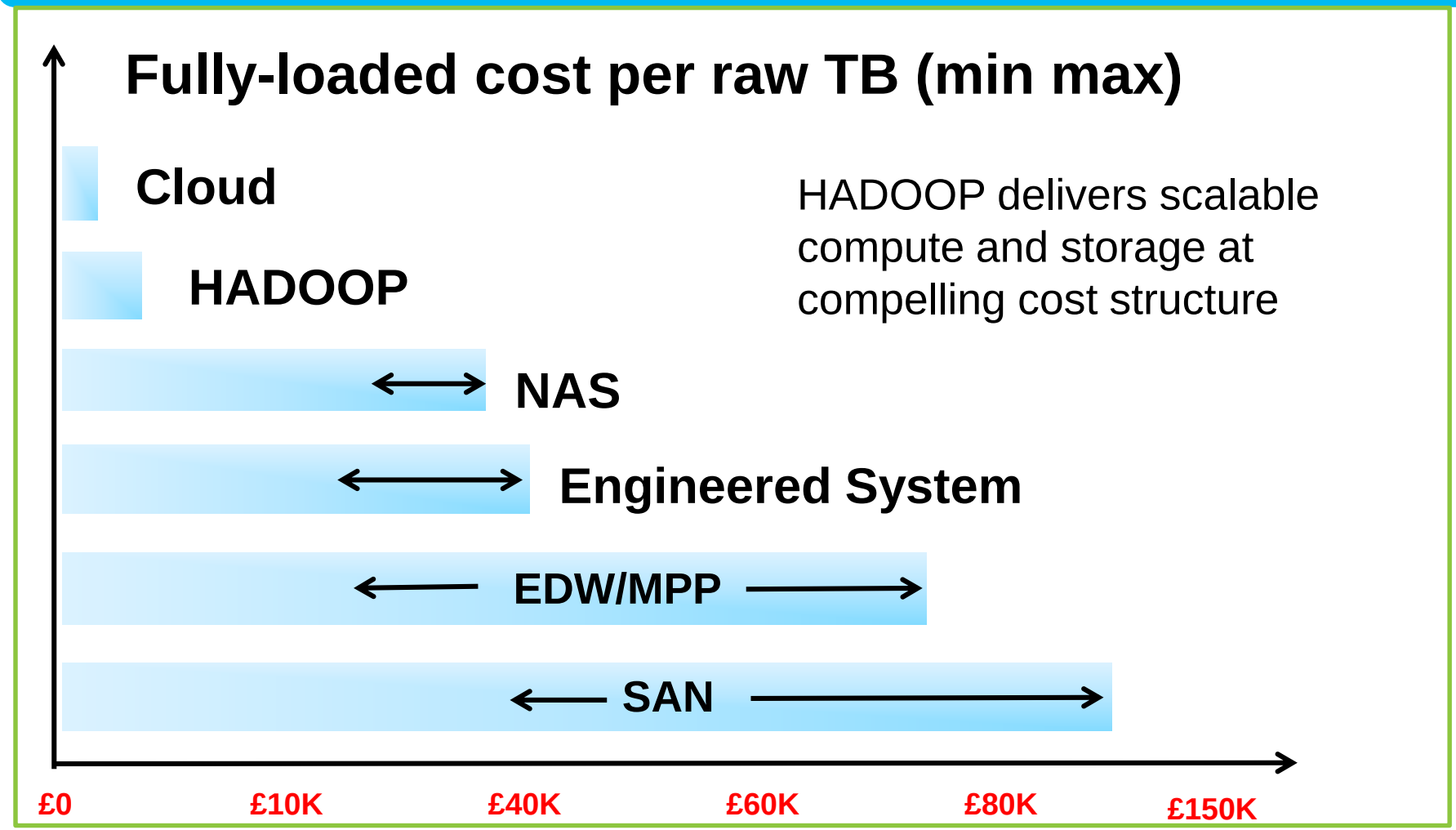
- *"Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs."*

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




© British Gas

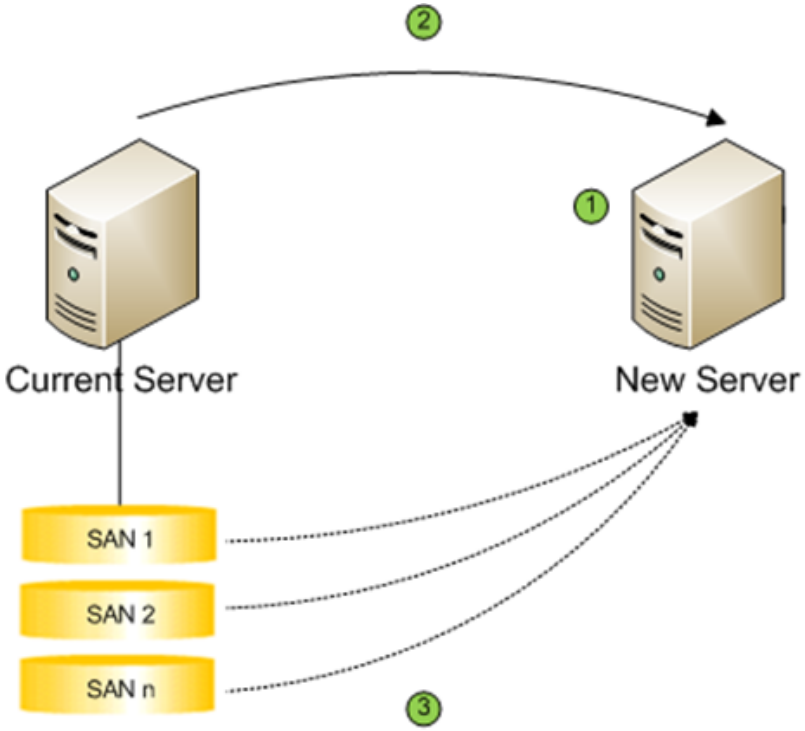
# Modern Data Architecture with commodity compute & storage




http://www.slideshare.net/fullscreen/hortonworks/distilling-hadoop-patterns-of-use-and-how-you-can-use-them-for-your-big-data-analytics/9

© British Gas

# Analytics on the Lake: Migration - setup approach

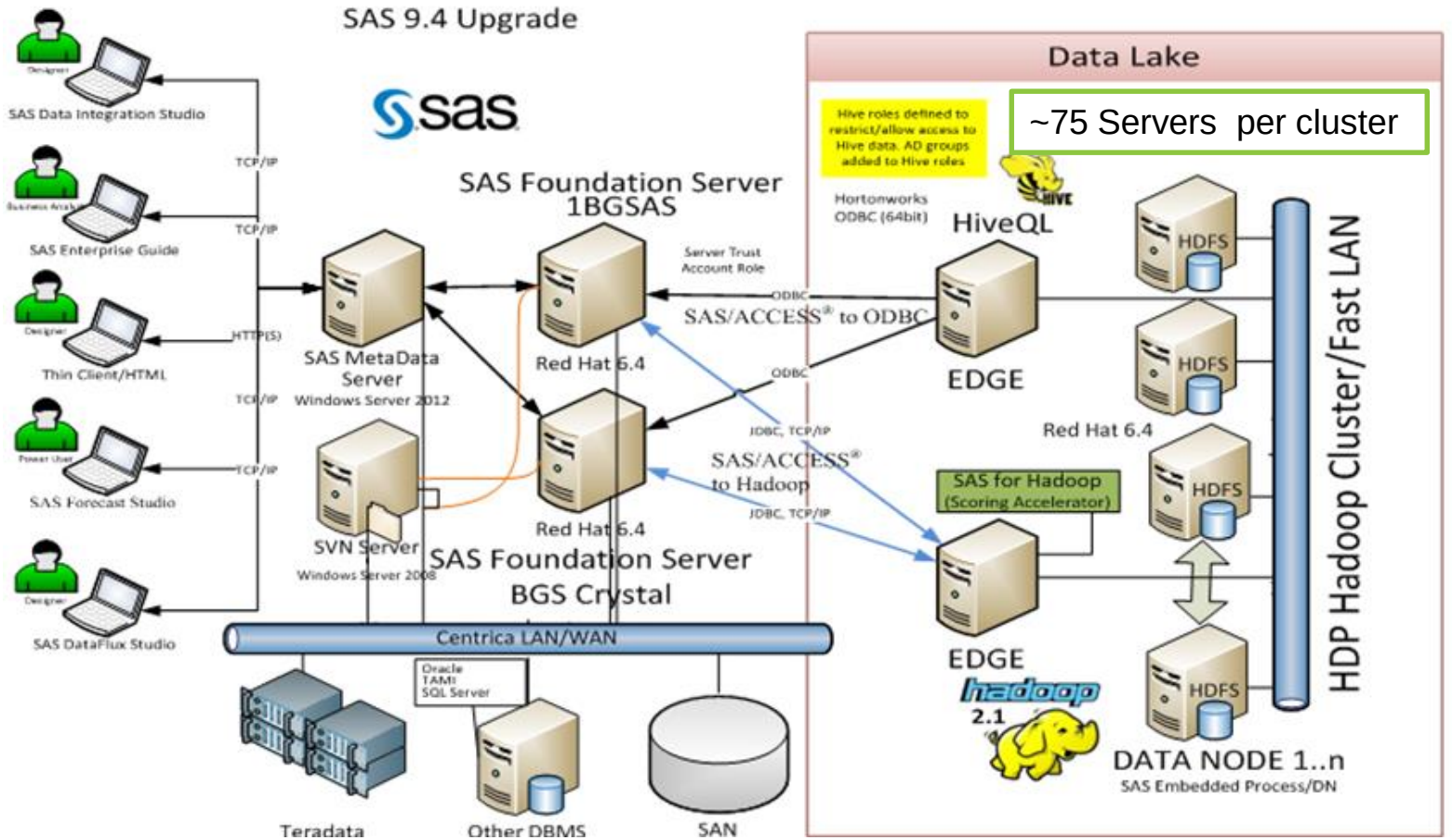



1. Build new Server with RHEL 6.4 and SAS 9.4
2. Migrate Team 1
3. Repoint Team 1 SAN

- **Analytics on the Lake Project**
  - Upgrade to SAS9.4 and connect to Hadoop
  - ~9 months to complete.
  - One Dev/Test Environment and Two Production Environments for 350 users.
  - Create new SAS9.4 instances and re-point storage to it.

© British Gas

# Our SAS 9.4 Environment

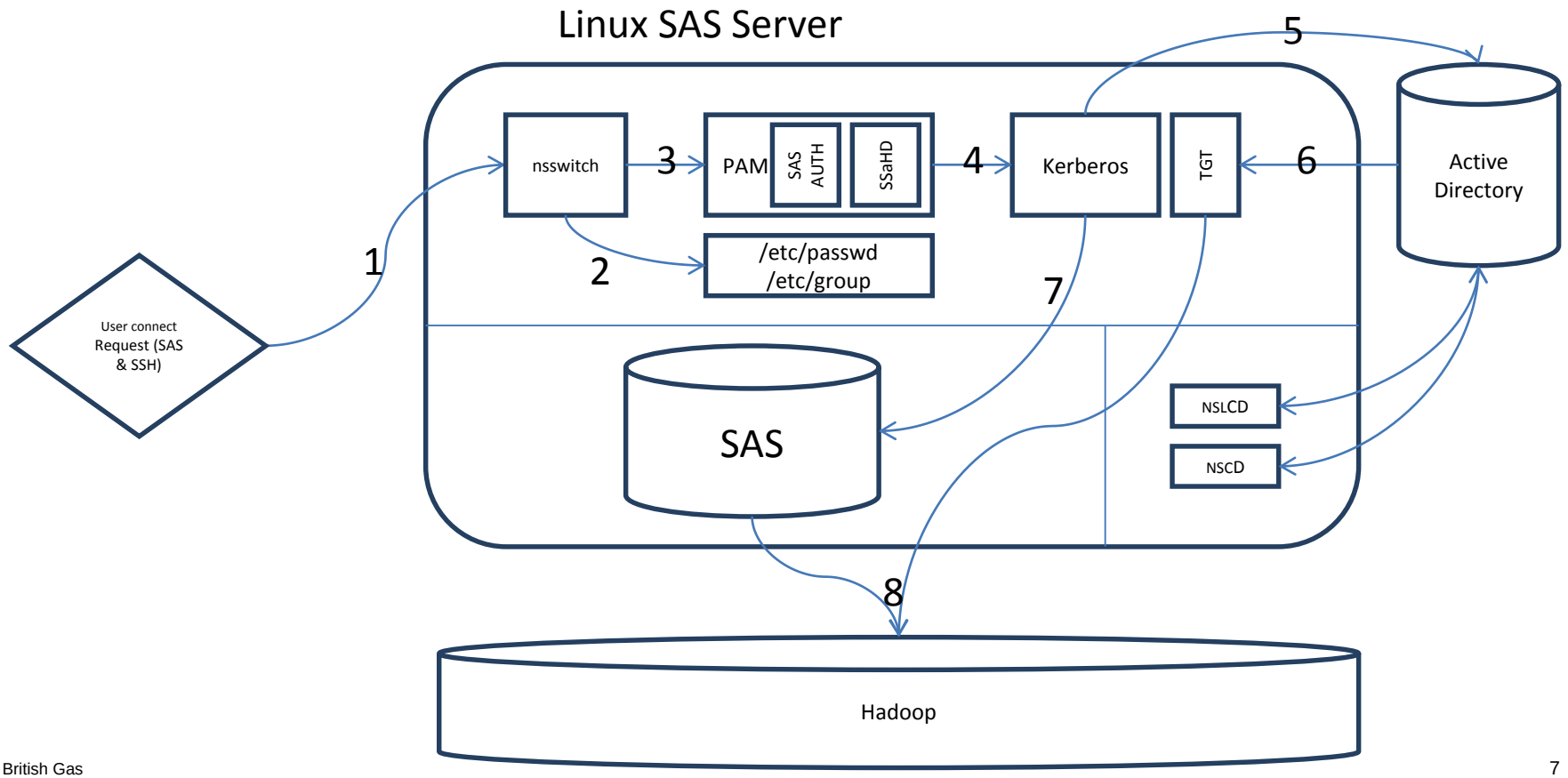



© British Gas

# Top Admin Issues

- Single Sign-On: Individual user Kerberos ticket required for Hadoop connection

→ **Solution:** Kerberos ticket provided by Active Directory, supplied by PAM (see appendix A for details)



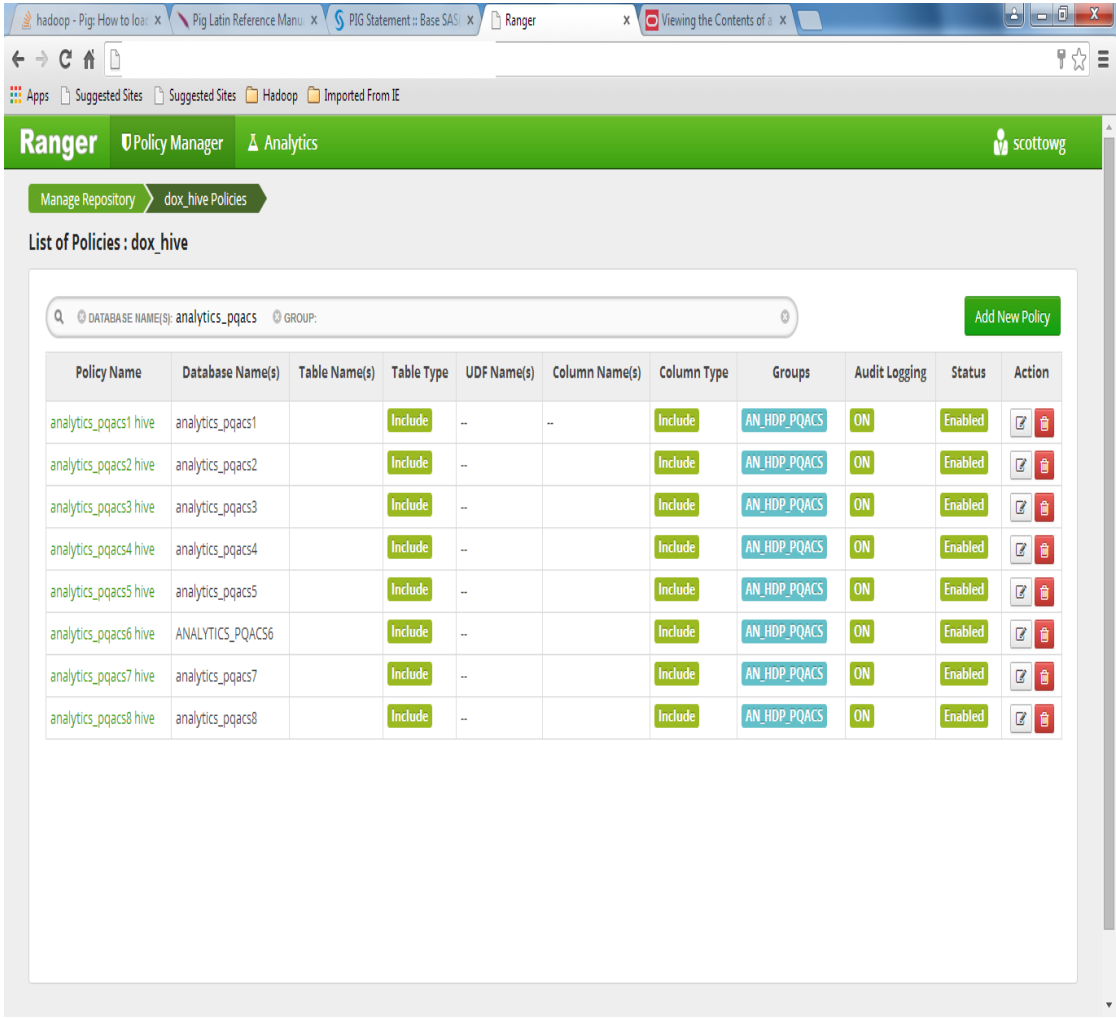

© British Gas

# Top Admin Issues

- No mechanism from SAS to administer / enforce security permissions on Hadoop

→ **Solution:** Apply Security / Permissions via Ranger

  - Access to Hadoop Database schemas is based on AD group membership.
  - We have matching AD groups in our metadata group membership.




© British Gas

# Top Admin Issues

- No default method to control storage allocation on Hadoop

→ **Solution**: Quotas set up on Hadoop Databases in conjunction with Hadoop Admins.

- Users have no visibility of storage allocation on Hadoop (no access to HDFS)

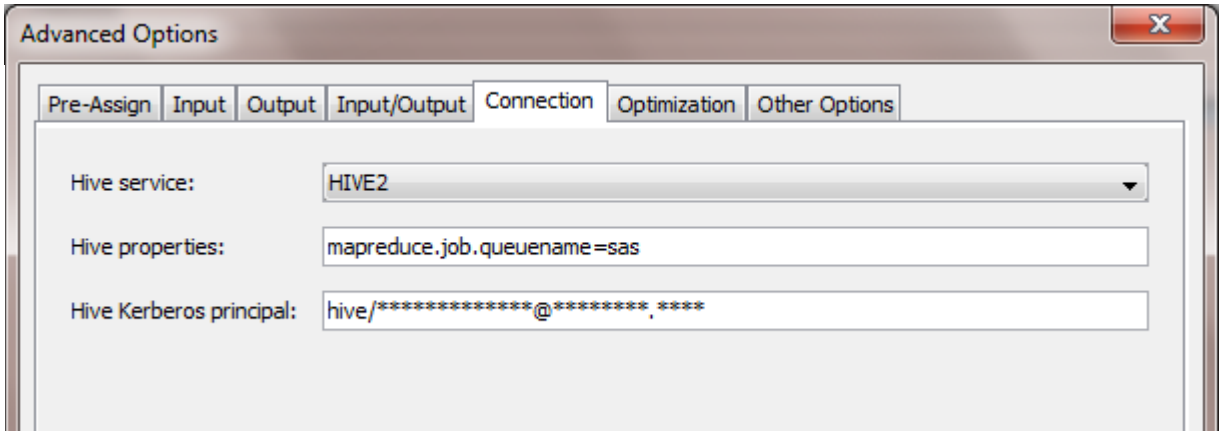
→**Solution**: *Tactical*: listing thru Pig shell in Hue fs -du /apps/hive/warehouse/myhivedb.db

*Strategic*: Prototyping an in-house SAS macro, utilising Proc Hadoop to execute a PIG script

- No resource prioritisation - users added to 'default' queue as standard

→ **Solution:** Hadoop queue set by Hadoop Admins to manage against list of users, users set job properties accordingly

•A dedicated SAS queue guarantees users 20% cluster resource with potential to extend to 30%
•All SAS library and pass-through definitions to Hadoop controlled by SAS admins (via Metadata and macros)

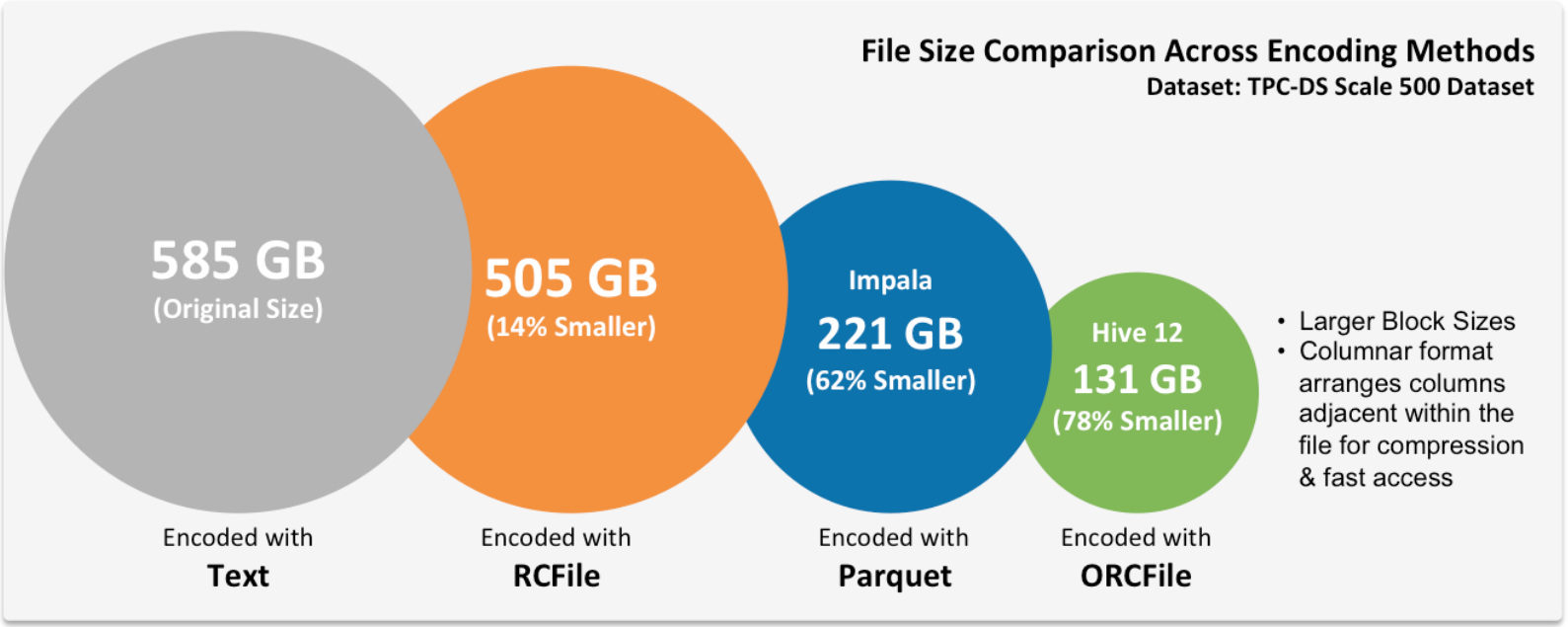



© British Gas

# Top Admin Issues

- Inefficient storage mechanism by default – Storage / optimisation
  - No mechanism to globally apply compression in Hadoop

→ **Solution:** The user is expected to select the file type they wish to use along with a selected compression. E.G. (ORC)Optimized Row Columnar using Snappy compression. Moving to ORC file format by default for its query performance.'




http://hortonworks.com/blog/orcfile-in-hdp-2-better-compression-better-performance/

© British Gas

# Admin Issues

- No standard mechanism to manage user connections

→ **Solution**: For Explicit pass though statements, a SAS Macro was created whereby the user passes the database as an input parameter

```
PROC SQL;
  CONNECT TO HADOOP as con1
   (%hdpconnect (Hadoop_schema));
```

- Hadoop default backup (Trash) stores 1 day's worth of deleted data

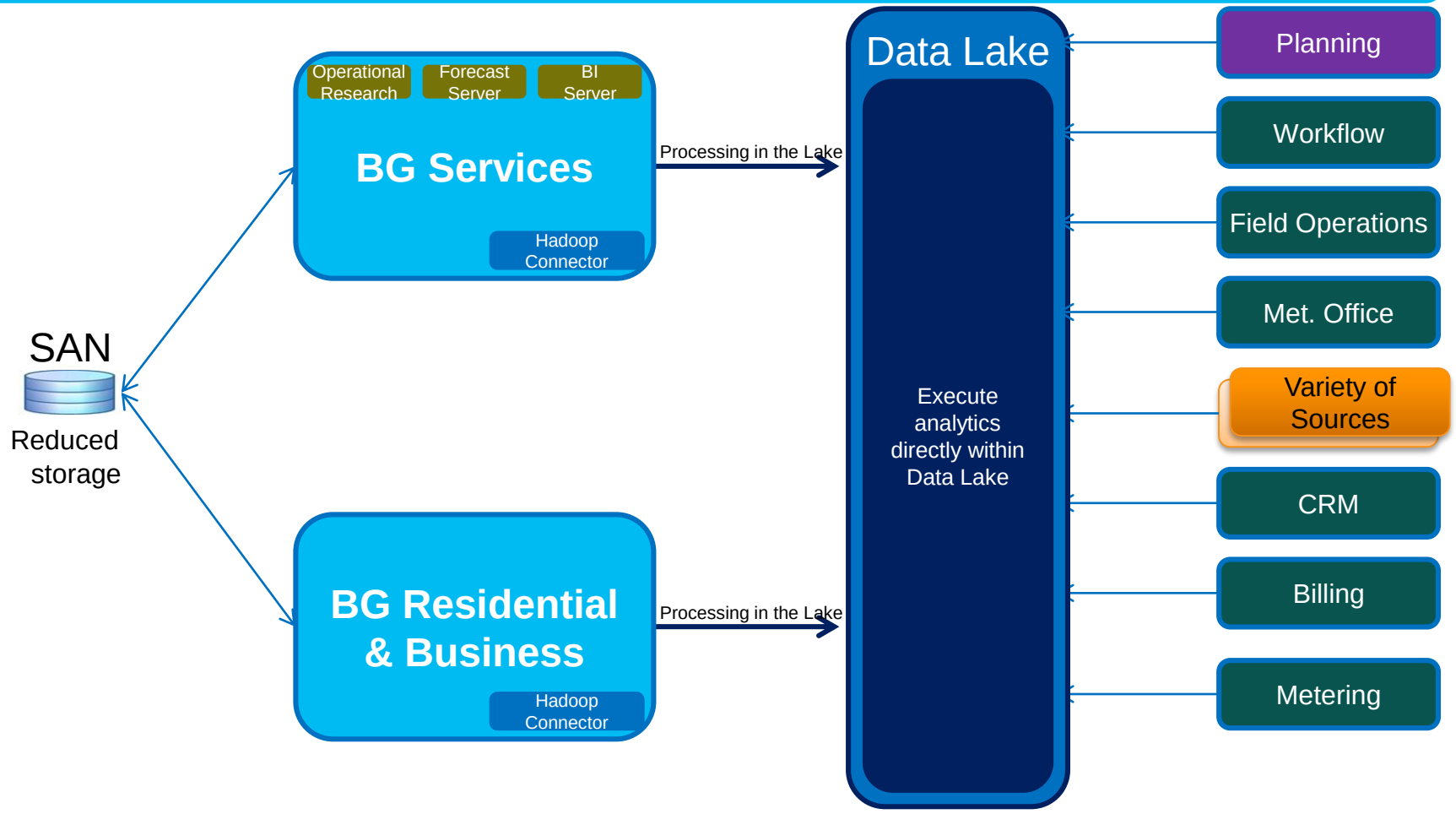
→ **Solution**: tactical solution provided to backup required files, detected using prefixes BUT the space to store this backup comes out of individual users' quota.

- No metadata solution provided with Hadoop Data Lake

→ **Solution:** Bespoke in-house tool created to manage lake metadata

© British Gas

# Admin Issues
# Multiple source systems → Single version of Truth




**Technology Legend:** ORACLE | SQL Server | Flat Files | Hadoop

© British Gas

# Top User Issues

- New methods of Coding for Users:
  - Data Step VS Implicit Passthru VS Explicit Passthru
  - Where is the code being executed? **Solution**: SAS Option for Trace Logging:

```
OPTIONS SASTRACE=',,,d' SASTRACELOC=SASLOG NOSTSUFFIX SQL_IP_TRACE=(note,
  source) msglevel=i;
```

```
HADOOP_1: Executed: on connection 4
  USE `hadoop_schema`
```

  - SAS WILL automatically overwrite datasets in libraries that a user has write access to, Hadoop WILL NOT – **Solution**: Code in to drop the Hadoop table first
  - Which way is most efficient? (Our Findings)

© British Gas

# Top User Issues

- New Skills Required – Explicit Passthru uses HiveQL

```
proc sql;
connect to HADOOP (
        %hdpconnect(hadoop_schema));
execute (
      create table class as
       select *
       from class2
       ) BY Hadoop;
disconnect from HADOOP;
quit;
```

This is HiveQL

- **Solution**: User Education in HiveQL

© British Gas

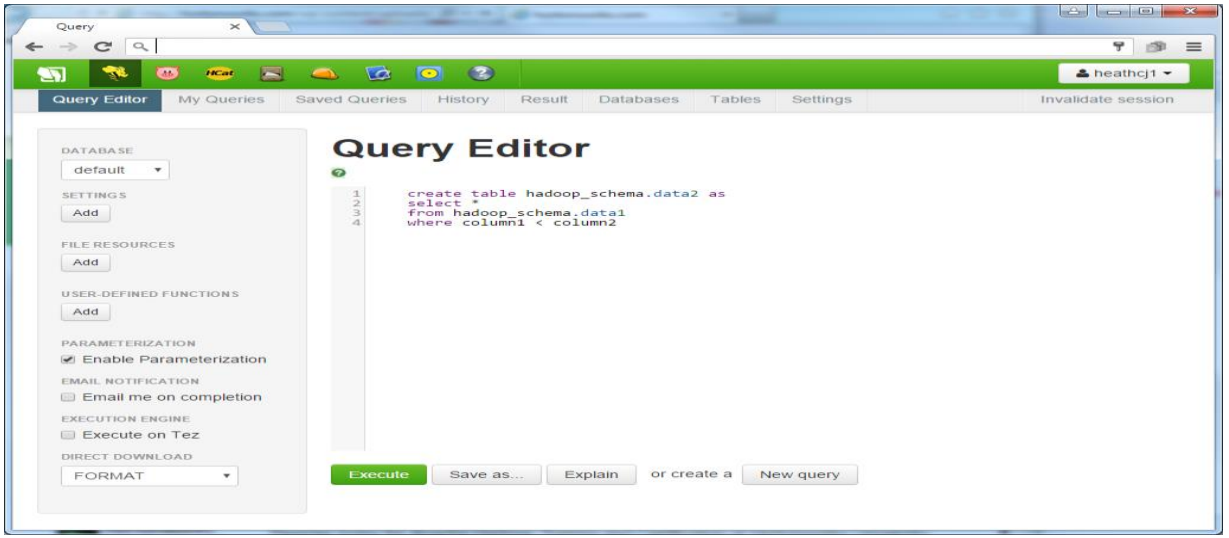
# Top User Issues

- Current Error Logging Limitations between SAS & Hadoop

ERROR: Unable to execute Hadoop query.

ERROR: Prepare error. SQL statement: create table hadoop_schema.data2 as select * from hadoop_schema.data1 where column1 < column2.

- **Solution**: Use Hue tool to circumvent




© British Gas

# Top User Issues

- String Format in Hadoop – Defaults to 32,767 Length in SAS

| Name | Type | Length | Format | Informat | Label |
|---|---|---|---|---|---|
| addmumber | Character | 32767 | $32767. | $32767. | addmumber |
| client | Character | 32767 | $32767. | $32767. | client |
| comm_type | Character | 32767 | $32767. | $32767. | comm_type |
| consnumber | Character | 32767 | $32767. | $32767. | consnumber |

- **Solution**: Create Casted Views in Hadoop

→ See appendix B for code example

| Name | Type | Length | Format | Informat | Label |
|---|---|---|---|---|---|
| addmumber | Character | 10 | $10. | $10. | addmumber |
| client | Character | 3 | $3. | $3. | client |
| comm_type | Character | 3 | $3. | $3. | comm_type |
| consnumber | Character | 3 | $3. | $3. | consnumber |

© British Gas

# Top User Issues

- EG 7.1 Wizard cannot use different Hadoop Databases

```
PROC SQL;
   CONNECT TO HADOOP as con1
    (%hdpconnect(hadoop_schema));
EXECUTE (
   CREATE TABLE SAS_EG_test AS
      SELECT "t1"."int_ui",
          "t1"."anlage",
          "t2"."ext_ui"
         FROM "ISU_EUITRANS" "t2"
           INNER JOIN "ISU_EUIINSTLN" "t1" ON ("t2"."int_ui" = "t1"."int_ui")) by con1;
   DISCONNECT FROM con1;
QUIT;
```

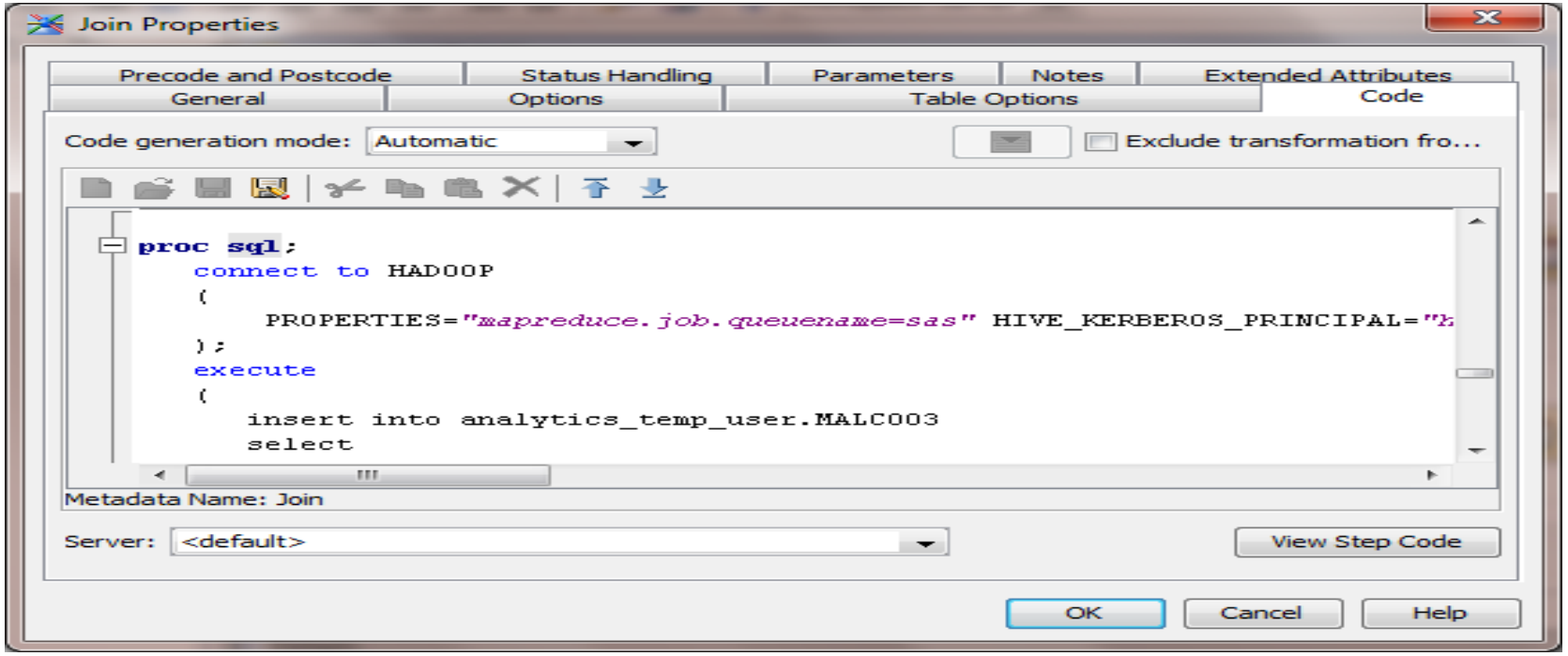
- **Solution**: Tactical workaround via code window

Longer term: hotfix / later version of EG

© British Gas

# Top User Issues

- Auto Generated Code from DI 4.9 (m2) Missing Word “table”




- **Solution**: Tactical using “User Written Body” – Fixed in DI Studio 4.9 (m3)

© British Gas

# Benefits / Success Stories - Speed

| Job Description | Previous System Duration | Hadoop Duration |
|---|---|---|
| Supply address information relating to contracts | 1 hr 50 min | 20 min |
| Open meter reading order and its age for the quarterly billed customers | 1 hr 30 min | < 30 min |
| Create a table of all the transactions from SAP in one simple to use table | Not possible due to Size | Between 2 and 3 hours |

© British Gas

# Top 5 Recommendations

- Upgrade quickly
- Close collaboration from the start between users, security and Hadoop Admin's
- Up front set-up of data governance process & security framework
- Empowerment of Users
- Data Ingestion can start now!!! (*data quality is key)*

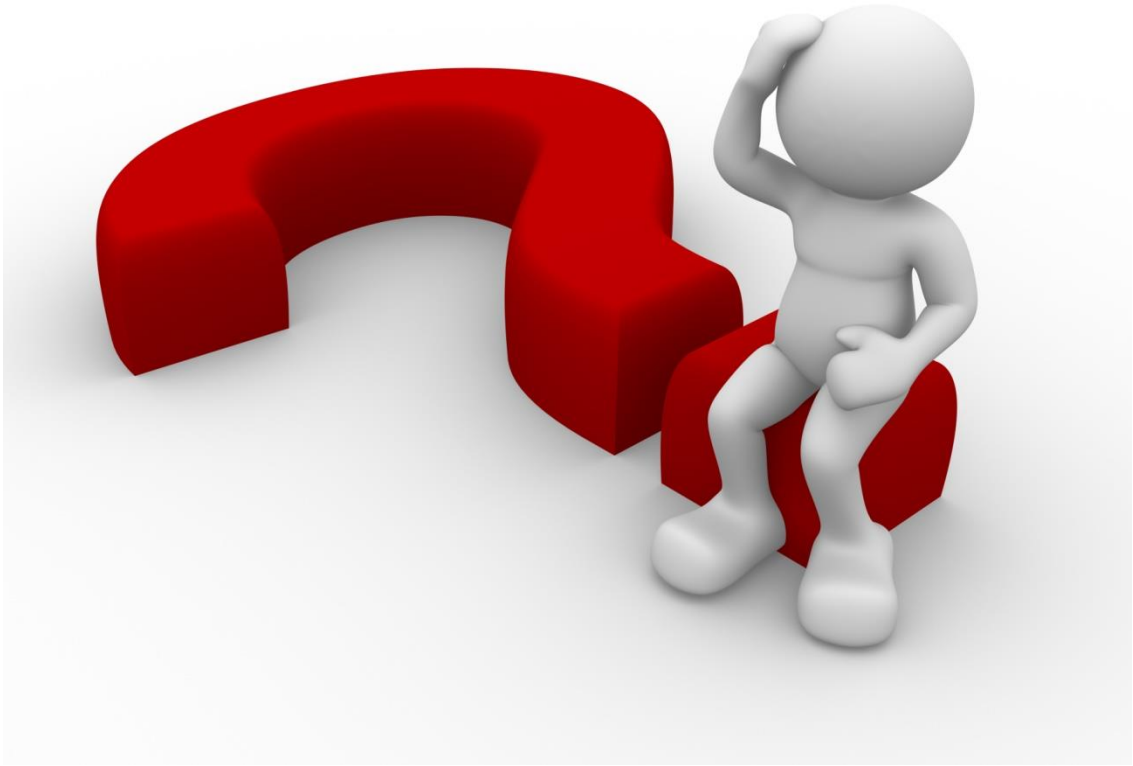
© British Gas

# Summary

- Utilisation of Hadoop Technology is a new and exciting opportunity
- Emerging SAS-Hadoop relationship
- Next Steps:
  - HDP v2.3
  - Migrating Storage to Hadoop
  - Realise Value with migration of Business Processes

© British Gas

# Questions



© British Gas

# APPENDIX A - Linux LDAP & Kerberos authentication model

- 1. User requests authentication to the Linux server via SAS client (e.g. Enterprise Guide).
- 2. The NSSwitch module dictates to the OS which authentication mechanism to use. In this case, local files first then LDAP (Active Directory).
- 3. If the user/password combination is not present in the local files, then progress to LDAP which is via PAM. If the request is not for SAS but instead Putty, then the SSHD module is referenced. If the request is for SAS then the SASAUTH module is referenced.
- 4. PAM is configured to use Kerberos to authenticate between the Linux machine and Active Directory (LDAP).
- 5. The Kerberos module handles the request for authentication against LDAP.
- 6. If authentication is successful, Kerberos will automatically generate a TGT (Ticket granting ticket) which is written as a temporary file on the Linux file system. This can be passed to Hadoop for authentication.
- 7. Once authenticated by Kerberos, the OS allows the SAS process to start as the AD user. An entry in the SAS configuration identifies the valid TGT file for the user and creates the necessary Hadoop environment variables.
- 8. SAS attempts to connect to Hadoop and automatically passes the TGT allowing seamless access to the data lake.
- The NSLCD module directly references LDAP to provide user/group information.
- The NSCD module provides caching of credentials to limit the queries sent to LDAP.

http://support.sas.com/documentation/installcenter/en/ikfdtnunxcg/66380/PDF/default/config.pdf

© British Gas

# APPENDIX B: Example of Casted Code

```
proc sql ;
connect to hadoop as hadoop (
  %hdpconnect (hadoop_schema));
execute by hadoop (
Create table hadoop_schema.hadoop_table_b as
select  cast(var_a as varchar(12)) as var_a,
        cast(var_b as bigint) as var_b,
        cast(CONCAT(SUBSTR(var_c,1,4),"-",SUBSTR(var_c,5,2),"-
  ",SUBSTR(var_c,7,2)) as date) as var_c
from hadoop_schema.hadoop_table_a
);
quit;
```

© British Gas