HadoopとSASとの連携テクニック

小林泉
SAS Institute Japan株式会社
ビジネス推進本部

Techniques in SAS on Hadoop

Izumi Kobayashi
Analytics Platform Practice, SAS Institute Japan

要旨：

ビッグデータ分析の基盤としてのHadoopの活用が進んでいます。SASはこのような時代の流れを考慮し、SASとHadoopを統合する製品を提供しています。本講演では、ビッグデータをSASで取り扱う機能とテクニックについて取り扱います。主にSASユーザーに向け、Hadoopクラスター上で動作するSASプログラミングテクニックをご紹介し、SASからHadoopの分散処理フレームワークをどのように活用できるかについて解説します。

キーワード： SAS, Hadoop, Map Reduce, Hive, Impala, HDFS, DS2, 機械学習, インメモリ, 分散処理

アジェンダ

- Hadoopとは
- SASにとってのHadoop
- SAS on Hadoop連携テクノロジー
  - Hadoopへのアクセス
  - Hadoop上でのSAS処理の実行
  - インメモリ処理

Hadoopとは

- 何のために開発された？

  ビッグデータを取り扱うためのデータ処理フレームワーク。

- どのような環境で使うの？

  ディスクやメモリを共有せずに、汎用サーバーを単純に非常に数多く並べた環境（クラスター）を想定。

- データはどのように配置されているの？

  データは、HDFS（Hadoop Distributed File System）を使用してクラスター上で複数サーバーにまたがって配置される。

- 処理はどのように行われるの？

  データは、 Hadoop上でのデータ処理フレームワークであるMapReduceと呼ばれる、Java言語ベースのプログラミングモデルを使用して処理される。

Hadoopが注目されている理由

- ビッグデータ
  - データの量の増大とデータ種類の多様化
  - 処理のスケーラビリティの必要性
- コモディティ化
  - 初期導入の敷居が下がる
  - 運用管理のしやすさ向上
  - SQLパフォーマンスの向上
- 経済性
  - 現実的なコストで実現したい

SASユーザーにとってのHadoop

- 新しいデータソース
  - 新しいデータベースとして
  - 新しいファイルシステムとして
- 分散並列処理基盤
  - SAS処理のパフォーマンス向上、スケーラビリティ向上
  - シングルプロセスから分散並列処理へ
- インメモリ分析基盤
  - SASの新しいインメモリ・アーキテクチャの基盤として

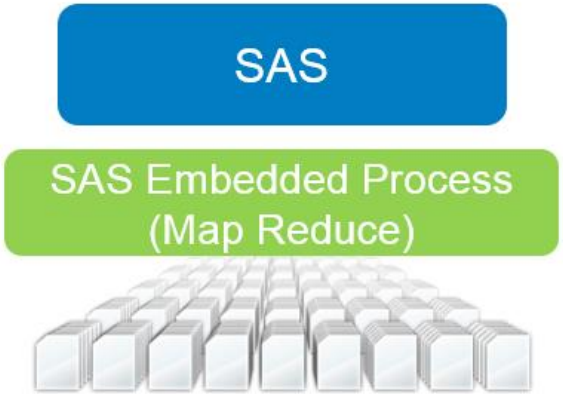
SAS & Hadoop　Hadoopとの統合を実現するSASのテクノロジー

FROM Hadoop
Hadoopデータへのアクセス

- Hive/Impala経由で処理を実行
- HadoopデータのRead/Write
- SASデータのHDFSへの配置

IN Hadoop
Hadoop内で実行

SASの定評のあるデータ加工処理(DS1&DS2)をHadoop内でMap Reduceとして実行



WITH Hadoop
インメモリアナリティクス

処理はインメモリで実行。Hadoopをデータストレージと汎用処理基盤として利用

SAS & Hadoopに必要なスキル

- 通常Hadoopで必要とされるJava言語スキルは不要
- 操作言語は従来と同じSAS言語
  - GUIツールもあり
- Hadoop上のデータにライブラリを定義
- バックグラウンドで、Hadoop形式の処理である、HiveQLや、Map/Reduceに自動的に変換されるため、それらの知識は不要

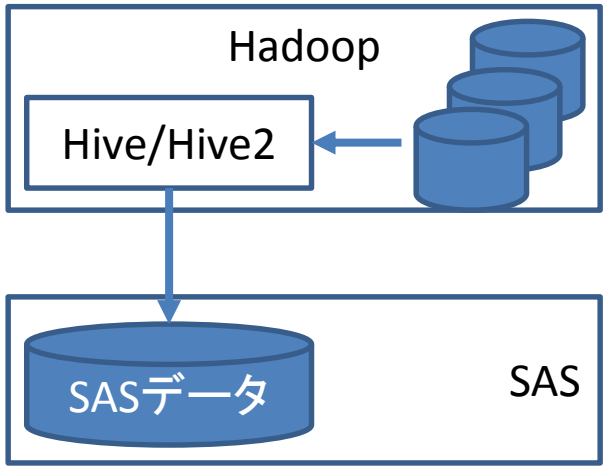
Hadoopデータにアクセス！

FROM Hadoop
Hadoopデータへのアクセス

SAS/Access to Hadoopを使用してHiveデータにアクセス

```
/* hadoopエンジンを使用してライブラリの定義 */
LIBNAME myhdp hadoop server=myserver user=sasdemo;

/* Hadoopからデータを抽出し、SASデータセットを作成 */
data work.mydata;
   myhdp.mydata;
   where actual >= 1000;
run;
```

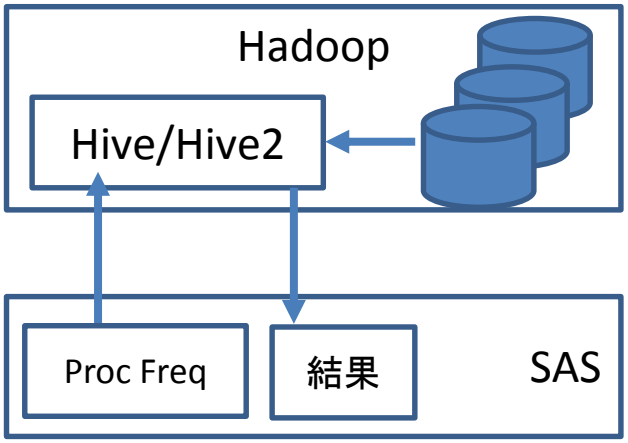


Hadoopデータにアクセス！

FROM Hadoop
Hadoopデータへのアクセス

```
/* Hadoop上にデータを作成 */
data myhdp.class;
   set sashelp.class;
run;

/* Hadoop上のデータに対して
      SASプロシジャを実行 */
proc freq data=myhdp.class;
   tables sex * age;
   where age >9;
run;
```



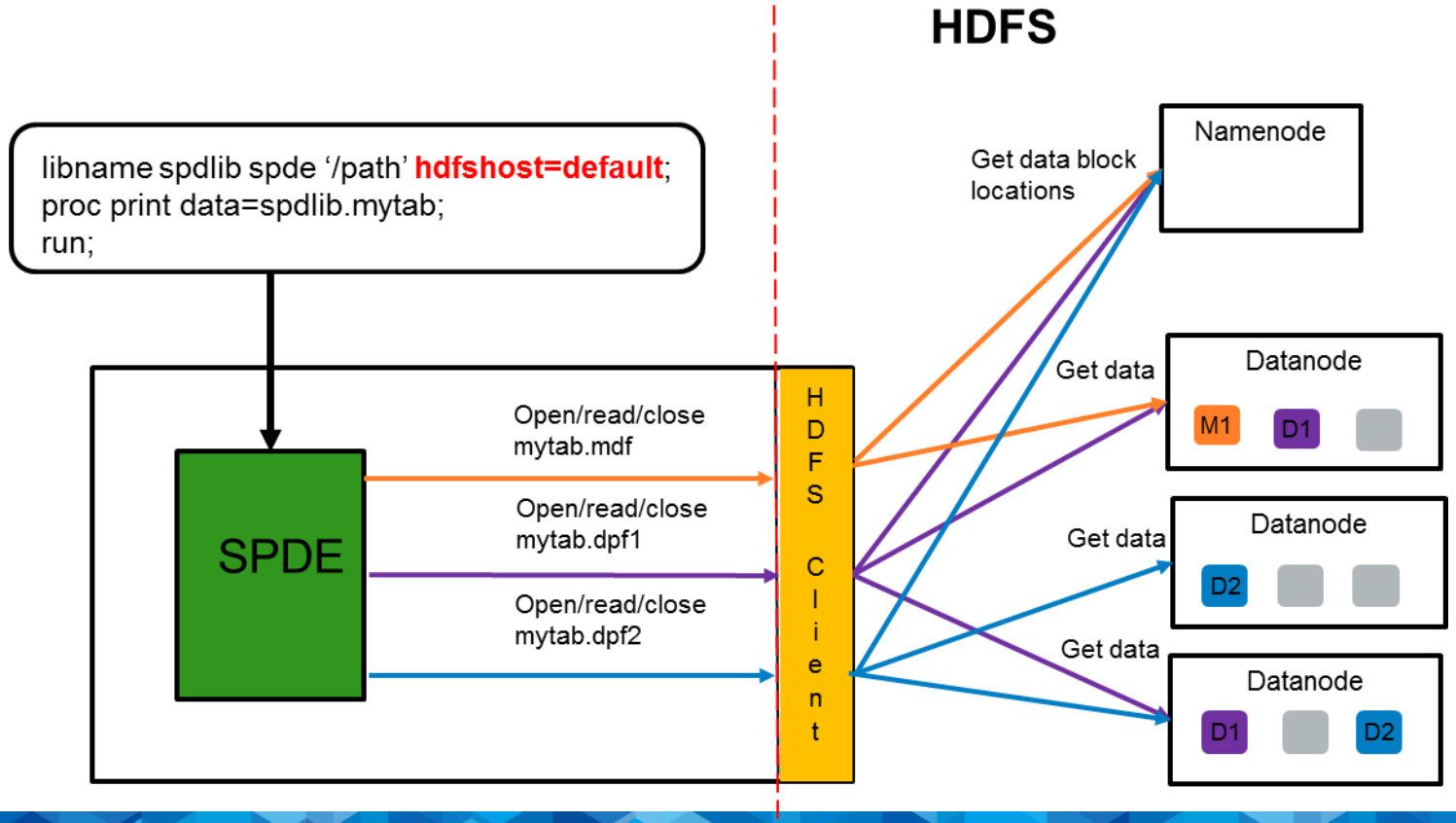

```
SQL_IP_TRACE: プッシュダウン数: 1
SQL_IP_TRACE: 伝達されたクエリ:    select
COUNT(*) as ZSQL1, case  when COUNT(*) >
COUNT(TXT_1.`sex`) then ' ' else
MIN(TXT_1.`sex`) end as
ZSQL2, case  when COUNT(*) >
COUNT(TXT_1.`age`) then  NULL  else
MIN(TXT_1.`age`) end as ZSQL3,
MAX(TXT_1.`age`) as ZSQL4 from `CLASS`
TXT_1 where TXT_1.`age` > 9 group by
TXT_1.`sex`, TXT_1.`age`
SQL_IP_TRACE: 一部の SQL は DBMS に直接渡されました。
NOTE: SQL生成を使用して度数表とクロス集計表を作成します。
```

★最新のアップデート
SAS9.4M3からPre-Productionながら、PROC TRANSPOSEも対応

SPDE on HDFS

FROM Hadoop
Hadoopデータへのアクセス

SPDEライブラリの参照先をHDFSにする



SPDE on HDFS

FROM Hadoop
Hadoopデータへのアクセス

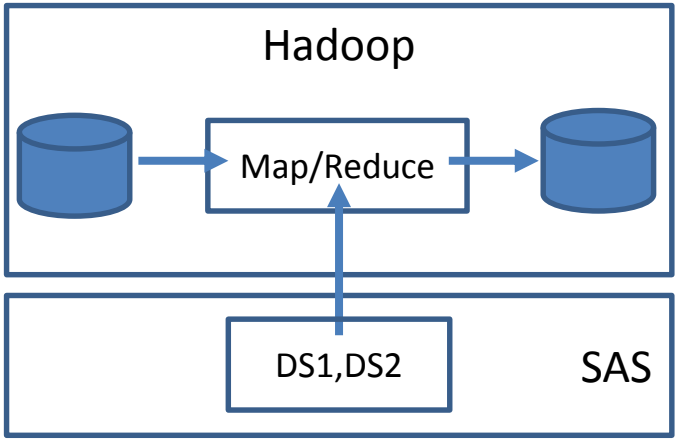
```
/* SPDE on HDFS */
libname hdfslib spde '/tmp/spde_hdfs/'hdfshost=default
                                    HDFSUSER=sasdemo partsize=16m;

/* HDFS上にSAS SPDEデータを作成 */
data hdfslib.class;
     set sashelp.class;
     do i=1 to 20000; output; end;
run;

/* HDFSコマンドで確認 */
hadoop fs -ls -R /tmp/spde_hdfs
drwxr-xr-x   - sasdemo supergroup          0 2014-06-30 20:40 /tmp/spde_hdfs/class
-rw-r--r--   3 sasdemo supergroup   16776192 2014-06-30 20:40 /tmp/spde_hdfs/class/class.dpf.00002c8f.0.1.spds9
-rw-r--r--   3 sasdemo supergroup    1463808 2014-06-30 20:40 /tmp/spde_hdfs/class/class.dpf.00002c8f.1.1.spds9
-rw-r--r--   3 sasdemo supergroup      43256 2014-06-30 20:40 /tmp/spde_hdfs/class.mdf.0.0.0.spds9
```

データステップ処理をHadoop上で実行

- データステップ処理がマルチスレッド化され、Map Reduce処理としてHadoop上で分散並列処理することが可能となった
- 従来型のデータステップDS1と、その拡張版であるDS2をサポート
- 全てのデータステップ機能がサポートされるわけではない



IN Hadoop
Hadoop内で実行

データステップをHadoop上で実行

```
36      data hdplib.myclass_out;
37      set hdplib.myclass;
38      run;

NOTE: Attempting to run DATA Step in Hadoop.
NOTE: Data Step code for the data set "HDPLIB.MYCLASS_OUT" was executed in the Hadoop
EP environment.
NOTE: DATA statement used (Total process time):
      real time           26.36 seconds
      cpu time            0.36 seconds

         Hadoop Job (HDP_JOB_ID), job_201406281704_0013, SAS Map/Reduce Job,
      http://jpnplimstat-cdh4-
00.jpn.sas.com:50030/jobdetails.jsp?jobid=job_201406281704_0013
         Hadoop Job (HDP_JOB_ID), job_201406281704_0013, SAS Map/Reduce Job,
http://jpnplimstat-cdh4-00.jpn.sas.com:50030/jobdetails.jsp?jobid=job_201406281704_0013
         Hadoop Version         User
         2.0.0-cdh4.6.0         hdfs

         Started At                     Finished At
         Jun 30, 2014 10:37:06 PM       Jun 30, 2014 10:37:27 PM
```

Map/Reduce処理が実行されているのが分かる

IN Hadoop
Hadoop内で実行

DS2をHadoop上で実行

```
proc ds2 indb=yes;
   thread tpgm / overwrite=yes;
   declare double highscore;
   keep gender highscore;
   retain highscore;
      method run();
         set hdptgt.myinputtable;
         by gender;
         if first.gender then highscore=0;
         if score1 <= 100 and score1>highscore then highscore=score1;
         if last.gender then output;
      end;
   endthread;
   run;
ds2_options trace;
   data hdptgt.myoutputscore(overwrite=yes);
      dcl thread tpgm tdata;
      method run();
         set from tdata;
      end;
   enddata;
   run;
quit;
```

By-Group処理も可能

★最新のアップデート
SAS9.4M3から、マージ（MERGE）をサポート

IN Hadoop
Hadoop内で実行

Hadoopデータをメモリ*にロード

* SAS® LASR Analytic Server

```
/* Hadoopクラスター上へライブラリを割り当て（HIVEを使わない方法）*/
libname hdp HADOOP
    config = &cfgFile
    hdfs_datadir = '/user/demo/files'
    hdfs_tempdir = '/user/demo/tmp'
    hdfs_metadir = '/user/demo/meta'
    hdfs_permdir = '/user/demo/perm';

/* HDFS上のCSVファイルのレイアウト定義をPROC HDMDを使用して作成 */
Proc HDMD name=hdp.weblog format=delimited encoding=utf8
          sep = ',' text_qualifier='"‘ DATA_FILE=‘weblog.csv';
     column ID int;
     column log varchar(1000);
run;
```

(次スライドへ続く)

WITH Hadoop
インメモリアナリティクス

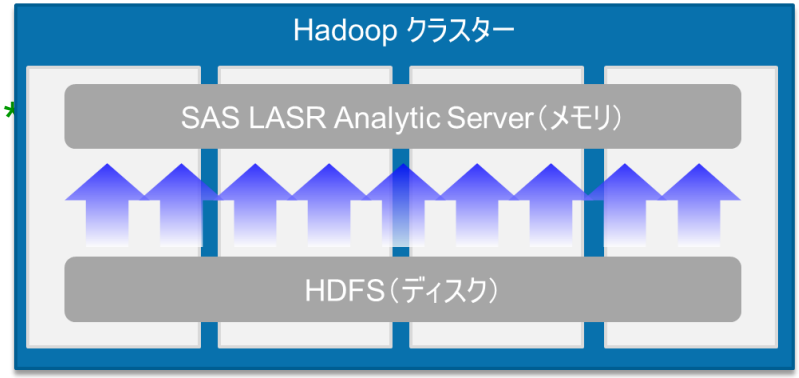
Hadoopデータをメモリ*にロード

（前スライドからの続き）

```
/* LASR Analytic Server を1314ポートで起動 */
options set=GRIDHOST="servername";
options set=GRIDINSTALLLOC="/opt/TKGrid";
proc lasr create port=1314
     path="/tmp" noclass;
     performance nodes=all;
run;

/* LASR Analytic Serverへデータをロード */
proc lasr add port=1314
               data=hdp.weblog;
run;

/* 1314ポートのLASRプロセスを終了 */
proc lasr term port=1314; run;
```



WITH Hadoop
インメモリアナリティクス

Hadoop上でのインメモリ分析

★最新のアップデート
SAS9.4M2から、グラフ理論を扱うHYPERGROUPステートメントを提供

```
/* LASRライブラリの定義 */
Libname LASRLIB SASIOLA SERVER=MYSERVER;

/* インメモリ処理用のプロシジャPROC IMSTATを利用 */
proc imstat data=LASRLIB.ASSOCS;
     arm /* Association Rule Mining アソシエーションルール ステートメント */
      item = PRODUCT  /* アイテムに該当する変数名を指定 */
      tran = CUSTOMER /* ユーザIDに該当する変数名を指定*/
   /* オプション */
   / minItems=2        /* minlen */
     maxItems=20        /* maxlen */
     support(lower=100) /* Support サポート値 */
     rules(confidence(lower=0.10))   /* confidence 信頼度 */
     rulesTbl;   /* ルールテーブルを一時テーブルへ出力 */
run;
quit;
```

WITH Hadoop
インメモリアナリティクス

SASのHadoop対応製品

- インタラクティブ分析ツール
  - 対話的
  - ビジュアル
  - インメモリ

⇒次スライド

- 高度モデリングツール
  - バッチ型
  - プロダクション・モデリング

⇒SAS® High-Performance Products

Hadoop上での対話型分析

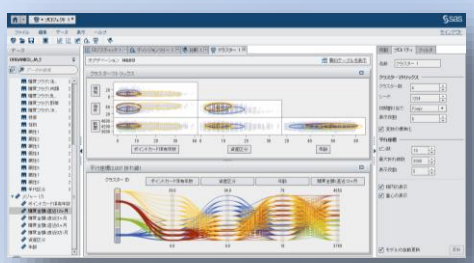
Hadoop上でのIn-Memory Analytics基盤（LASR Analytic Server）
同一の基盤、同一のデータを全てのユーザーで共有

グラフィカル・インターフェース

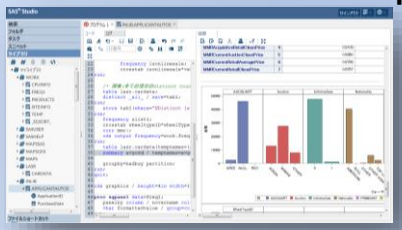
インタラクティブ・プログラミング

SAS® Visual Analytics

SAS® Visual Statistics



SAS® In-Memory Statistics for Hadoop



レポート・可視化・探索

統計解析・機械学習・予測モデリング

WITH Hadoop
インメモリアナリティクス

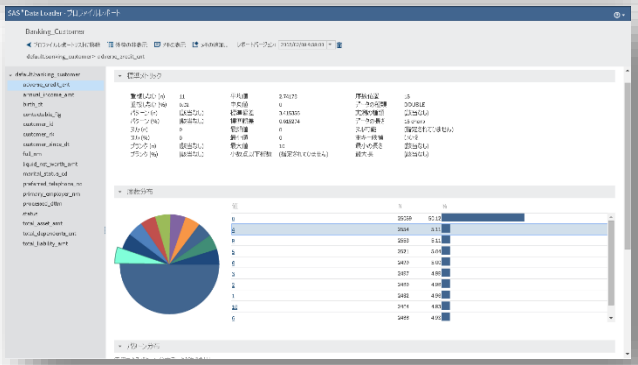
Hadoop上での対話型分析

SAS® Data Loader for Hadoop

ポイント&クリックによるHadoop上のデータの簡易な管理・加工

処理のタイプに応じて異なるタイプの処理コードを自動的に生成

- SAS DS2コード
- SAS Data Quality処理
- Hiveクエリー
- Sqoop & Oozie



IN Hadoop
Hadoop内で実行

参考文献

- SAS® 9.4 Support for Hadoopテクニカルサポートサイト
  http://support.sas.com/resources/thirdpartysupport/v94/hadoop/index.html
  - Understanding How SAS and Hadoop Work Together
  - Deployment: Installation and Administration
  - Support Resources from SAS: Communities, Blogs, Professional Services, and Training
- SAS/Access to Hadoop Reference
  http://support.sas.com/documentation/cdl/en/acreldb/68028/HTML/default/viewer.htm#n0kgg6z8c14ewmn1phdwdm5cp51i.htm
- SAS LASR Analytic Server Reference
  http://support.sas.com/documentation/onlinedoc/lasrserver/index.html
- DS2 Reference
  http://support.sas.com/documentation/cdl/en/ds2ref/68052/HTML/default/viewer.htm
- Introduction to SAS® Data Loader: The Power of Data Transformation in Hadoop
  http://support.sas.com/resources/papers/proceedings15/SAS1845-2015.pdf
- SAS and Hadoop
  http://support.sas.com/resources/papers/proceedings15/SAS4642-2015.pdf
- DS2 with Both Hands on the Wheel
  http://support.sas.com/resources/papers/proceedings15/2523-2015.pdf
- Tips and Techniques for Efficiently Updating and Loading Data into SAS® Visual Analytics
  http://support.sas.com/resources/papers/proceedings15/SAS1905-2015.pdf
- The SAS® Scalable Performance Data Engine: Moving Your Data to Hadoop without Giving Up the SAS Features You Depend On
  - http://support.sas.com/resources/papers/proceedings15/SAS1956-2015.pdf