Big Data Analytics met Hadoop

Jos van Dongen
Arno Klijnman
What is...

Hadoop

- HDFS
- Map/Reduce

Distributed storage and processing of (big) data on large clusters of commodity hardware
HDFS - Distributed storage for big files
Map/Reduce - Distributed processing for **big** data
The Hadoop Jungle
SAS & Hadoop Capabilities

WITH Hadoop

ON Hadoop

IN Hadoop

• SAS Data Quality Accelerator
• SAS Scoring Accelerator
• SAS Code Accelerator
SAS & Hadoop Integration

- User Interface
- Metadata
- Data Access
- Data Processing
- File System

- SAS® Data Management
- SAS® Enterprise Miner™
- SAS® Studio
- SAS® Visual Analytics
- SAS® In-memory Statistics for Hadoop

- SAS Metadata
- In-Memory Data Access
- Base SAS & SAS/ACCESS® to Hadoop™
- In-Memory Data Access

- Pig
- Hive
- Map Reduce
- SAS Embedded Process

- HDFS
- SAS® LASR™ Analytic Server
- SAS® LASR™ Server
Two Paradigms

Hadoop as a Data Platform

Hadoop as a core component of next generation analytical platform
Paradigm two

Hadoop as a core component of next generation analytical platform

- SAS/ACCESS
- SAS Data Management
- SAS Federation Server
- SAS Event Stream Processing
- SAS Data Loader for Hadoop
  - SAS Data Quality Accelerator for Hadoop
  - SAS Code Accelerator for Hadoop

- SAS Scoring Accelerator for Hadoop
- SAS Decision Manager
- SAS Visual Analytics

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

- SAS High Performance Analytics Products
- SAS Visual Statistics
- SAS In-memory Statistics for Hadoop
SAS runs the Entire Analytical Lifecycle in/on/with Hadoop

- SAS Visual Analytics
- SAS Scoring Accelerator for Hadoop
- SAS Code Accelerator for Hadoop
- SAS High Performance Analytics Offerings
  - SAS In-Memory Statistics for Hadoop
- SAS Visual Statistics
- Done using either the Data Preparation, Data Exploration or Build Model Tools

- BASE SAS
- SAS / Access
- SAS Data Loader for Hadoop
- SAS DI Studio
- SAS High Performance Analytics Offerings
- SAS In-Memory Statistics for Hadoop
- SAS Visual Statistics
- Done using the Build Model Tools and other checks

Copyright © 2014, SAS Institute Inc. All rights reserved.
USER ROLES & THE ANALYTICS LIFECYCLE

**BUSINESS MANAGER**
- Domain Expert
- Makes Decisions
- Evaluates Processes and ROI

**BUSINESS ANALYST**
- Data Exploration
- Data Visualization
- Report Creation

**IT SYSTEMS / MANAGEMENT**
- Model Validation
- Model Deployment
- Data Preparation

**ANALYST DATA SCIENTIST**
- Exploratory Analysis
- Descriptive Segmentation
- Predictive Modeling

**Domain Expert**
- Makes Decisions
- Evaluates Processes and ROI

**BUSINESS**
- Manager

**DATA**
- Preparation
- Exploration
- Visualization
- Report Creation

**IDENTIFY / FORMULATE PROBLEM**

**TRANSFORM & SELECT**

**BUILD MODEL**

**VALIDATE MODEL**

**DEPLOY MODEL**

**EVALUATE / MONITOR RESULTS**
USER ROLES & THE ANALYTICS LIFECYCLE

BUSINESS MANAGER
- Domain Expert
- Makes Decisions
- Evaluates Processes and ROI

DATA PREPARATION
- Model Validation
- Model Deployment
- Data Preparation

DATA EXPLORATION
- Exploratory Analysis
- Descriptive Segmentation
- Predictive Modeling

IDENTIFY / FORMULATE PROBLEM

BUILD MODEL

TRANSFORM & SELECT

VALIDATE MODEL

DEPLOY MODEL

EVALUATE / MONITOR RESULTS

BUSINESS ANALYST
- Data Exploration
- Data Visualization
- Report Creation

IT SYSTEMS / MANAGEMENT

ANALYST DATA SCIENTIST
- Exploratory Analysis
- Descriptive Segmentation
- Predictive Modeling
USER ROLES & THE ANALYTICS LIFECYCLE

BUSINESS MANAGER
Domain Expert
Makes Decisions
Evaluates Processes and ROI

BUSINESS ANALYST
Data Exploration
Data Visualization
Report Creation

IT SYSTEMS / MANAGEMENT
Model Validation
Model Deployment
Data Preparation

ANALYST DATA SCIENTIST
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

IDENTIFY / FORMULATE PROBLEM
DATA PREPARATION
DATA EXPLORATION
TRANSFORM & SELECT
BUILD MODEL
VALIDATE MODEL
DEPLOY MODEL
EVALUATE / MONITOR RESULTS

Domain Expert Makes Decisions Evaluates Processes and ROI

DEPLOY MODEL
VALIDATE MODEL
BUILD MODEL
TRANSFORM & SELECT
BUILD MODEL
DATA EXPLORATION
DATA PREPARATION
EVALUATE / MONITOR RESULTS
IDENTIFY / FORMULATE PROBLEM

Model Validation Model Deployment Data Preparation

Exploratory Analysis Descriptive Segmentation Predictive Modeling

SAS®Forum Nederland 2014
SAS® Data Loader for Hadoop
A new SAS Web-based Business user interface

What directive do you want to perform?

- **Saved Directives**: Open a previously created directive to run, view or edit.
- **Run Status**: Show the status of current and previous directive executions.
- **Query a Table in Hadoop**: Create a table by querying the data in an existing Hadoop table.
- **Run a SAS Program**: Run in-database data quality SAS programs.
- **Transform Data In Hadoop**: Transform data from a Hadoop table.
- **Load Data to LASR**: Copy data from a source and load it into LASR. Existing data in the target table will be replaced.
- **Profile Data**: Generate a profile report of the data in a table.
- **Saved Profile Reports**: Explore previously generated profile reports.

Enables Self-Service approach to managing data in Hadoop environment
SAS® Data Loader for Hadoop
Transform Data in Hadoop

Filtering Rules
- Source Table: Default > csv
- Manage Columns: 5 of 5 Columns
- Filter Data: year > 1900 | sales > 5

Column Selections
- Select the columns you want to include in the target data file
- Available columns: sales, year, month, day, date
- Selected columns: year, month, day, date

Aggregation
- Select the rows you want to summaries
- Group rows by: month
- Add Column
- Summary column: sales
- Aggregation: Sum
- New column name: SASA_SUM

No coding, scripting or specialized skills required
SAS® Data Loader for Hadoop

Query Hadoop data

Select Source Tables

Apply Query Criteria

See subset of data in Table Viewer

Simple Drag & Drop approach to Query Data inside Hadoop
SAS® Data Loader for Hadoop
Profile Hadoop Data

Select Source Table

Choose the data source with the data you want to query:

- default: Default Hive database
- downdwpx_01: DL dev schema
- downdwpx_01: DL dev schema
- downdwpx02: DC dev schema 2
- downdwpx01: Data Quality - Schema 1
- downdwpx_common: Common shared tables containing all common tables to use for testing
- downdwpx_common_test: Test tables containing all common tables to use for testing
- downdwpx_espec: ES tables for testing
- downdwpx_espec_directive: ES tables for testing

Run standard metrics on data inside Hadoop and generate reports

View Reports in Column Display

View Reports in Table Display
**Standard Metrics**

- **Ordinal Position**: 4
- **Count**: 1688948
- **Null (n)**: 0
- **Null (%)**: 0
- **Blank (n)**: 0
- **Blank (%)**: 0
- **Min. Value**: F
- **Max. Value**: U
- **Mode**: F
- **Unique (n)**: 3
- **Unique (%)**: 0
- **Pattern (n)**: 1
- **Pattern (%)**: 0.0
- **P.K. Candidate**: no
- **Data Type**: String
- **Data Length**: 1 chars
- **Actual Type**: string
- **Min. Length**: 1
- **Max. Length**: 1
- **Mean**: (not applicable)
- **Median**: (not applicable)
- **Nullable**: (not specified)
- **Decimal Places**: (not specified)
- **S. D.**: (not applicable)
- **S. E.**: (not applicable)

**Frequency Distribution**

<table>
<thead>
<tr>
<th>Value</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>923324</td>
<td>54.67</td>
</tr>
<tr>
<td>M</td>
<td>441940</td>
<td>26.17</td>
</tr>
<tr>
<td>U</td>
<td>323684</td>
<td>19.16</td>
</tr>
</tbody>
</table>
Query a Table in Hadoop

SOURCE TABLE  default / organics

SUMMARIZE ROWS  Group by: demgender, targetbuy / purchase_12mon: Sum

Select the columns to group by and the columns to summarize within the groups

Group rows by:
- demgender
- targetbuy

Summarize column:
- purchase_12mon

Aggregation:
- Sum

New column name:
- total_purchase_12mon

Next

FILTER ROWS  All rows

COLUMNS  Columns specified in “Summarize Rows”

SORT  (none)

TARGET TABLE  default / totals_organic

CODE  (generated code)

RESULT  Successfully queried data
<table>
<thead>
<tr>
<th>demgender</th>
<th>targetbuy</th>
<th>total_purchase_12mon</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Bought</td>
<td>1339040186.7499948</td>
</tr>
<tr>
<td>F</td>
<td>No</td>
<td>2529691969.1399984</td>
</tr>
<tr>
<td>M</td>
<td>Bought</td>
<td>307640939.4100002</td>
</tr>
<tr>
<td>M</td>
<td>No</td>
<td>1544105482.4699843</td>
</tr>
<tr>
<td>U</td>
<td>Bought</td>
<td>106367230.57000038</td>
</tr>
<tr>
<td>U</td>
<td>No</td>
<td>1249837525.04001</td>
</tr>
</tbody>
</table>
SAS® Data Loader for Hadoop
Copy Data to distributed SAS® LASR server

Select Source Table
Copy Data To distributed SAS® LASR Servers

Visualize Data

Explore Hadoop data quickly and easily for faster insights
USER ROLES & THE ANALYTICS LIFECYCLE

IDENTIFY / FORMULATE PROBLEM
DATA PREPARATION
DATA EXPLORATION
TRANSFORM & SELECT
BUILD MODEL
VALIDATE MODEL
DEPLOY MODEL
EVALUATE / MONITOR RESULTS

BUSINESS MANAGER
Domain Expert
Makes Decisions
Evaluates Processes and ROI

BUSINESS ANALYST
Data Exploration
Data Visualization
Report Creation

IT SYSTEMS / MANAGEMENT
Model Validation
Model Deployment
Data Preparation

ANALYST
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

DATA SCIENTIST
USER ROLES & THE ANALYTICS LIFECYCLE

**Business Manager**
- Domain Expert
- Makes Decisions
- Evaluates Processes and ROI

**Business Analyst**
- Data Exploration
- Data Visualization
- Report Creation

**IT Systems / Management**
- Model Validation
- Model Deployment
- Data Preparation

**Analyst**
- Exploratory Analysis
- Descriptive Segmentation
- Predictive Modeling

**Data Scientist**
- Build Model
- Transform & Select
- Validate Model

**Deploy Model**
- Evaluate / Formulate Problem
- Identify / Monitor Results

**Data Preparation**
- Data Exploration
- Data Visualization

**Results**
- Model Validation
- Model Deployment
- Data Preparation

**Analyst**
- Exploratory Analysis
- Descriptive Segmentation
- Predictive Modeling

**Data Scientist**
- Build Model
- Transform & Select
- Validate Model

**Deploy Model**
- Evaluate / Formulate Problem
- Identify / Monitor Results

**Results**
- Model Validation
- Model Deployment
- Data Preparation
SAS Scoring Accelerator for Hadoop
**** Have the SAS proc genmod run a logistic regression model ****

---

SAS Code Generated by LASS Analytic Server

Date: 29Sep2014 08:19:31
Model Type: Logistic Regression
Class variable: DemGender
Class variable: DemAge
Class variable: DemHomeowner
Class variable: PromClass
Response variable: TargetBuy
Distribution: Binary
Link Function: Logit

---

/** Defining temporary arrays and variables */
array xrow_0(19) temporary_
array beta_0(19) temporary_ 0
0.273482642616585
-0.0236949717141
1.97324667178033
0.996802178251566
0
-0.070572408305064
-0.20582048166369
-0.135497594141
-0.1395926803968
0.6694562962555
0
0.008370889260452
0
-0.02034794541427
-0.31929029339222
-0.0917008183943
0);

drop_badval__line__temp__i:_
badval=0;
line=0;
temp=0;
## Demo flow

<table>
<thead>
<tr>
<th>SAS DI</th>
<th>Data Loader</th>
<th>VA Explorer</th>
<th>VS</th>
<th>IMSTAT</th>
<th>Scoring Accelerator</th>
</tr>
</thead>
</table>
| • Access to Hadoop  
• Transform  
• Write back to Hadoop  
• Write to LASR | • Show table  
• Profile  
• Build Query  
• Write result to LASR | • Discover relations  
• Understand the data | • Discover a model  
• Determine significance  
• Cluster variables | • Recommendation  
• Dataset to enrich original dataset with recommendation results  
• Write to LASR | • Deploy model  
• Run model  
• Back to Data Loader |

- **SAS Data Management**
- **SAS Interactive Analytics On Hadoop**
- **SAS Analytics**
SAS & Hadoop: 3 Things to Remember

WITH Hadoop

ON Hadoop

IN Hadoop
Elastic IP Address

Setup:
- CentOS operating system
- Local users on all Amazon servers
- Internal network for all Amazon Servers
- Open firewall for all ports between workstation & server
- No integration Mail server
- No SSL

Internet

Demo Environment Infrastructure
SAS® Solutions for Hadoop

Hadoop brings big data. We bring everything else.

View video
More value from Big Data with Hadoop and SAS

9 oktober 2014
Huizen

Greater analyst productivity and power to solve complex problems by uncovering undetected patterns and trends faster than ever with SAS and Hadoop.

SAS software for Hadoop enables multiple users to concurrently manage and prepare data stored in Hadoop, explore and visualize this data, develop accurate statistical and machine learning models quickly, as well as access, deploy and execute these models in their Hadoop ecosystem.

SAS empowers organizations to start exploring the value of big data in Hadoop rather than just collecting and storing the data. Statisticians and Data Scientists can address all steps of the analytical lifecycle, going from raw Hadoop data to integrating analytical models into business processes. By using state-of-the-art statistical algorithms and machine-learning techniques, multiple users can concurrently explore and use multiple analytic approaches to build models and quickly run multiple iterations to determine the best models. This vastly improves end-user productivity of scarce data scientist resources, increases model building efficiency, decreases time from model inception to deployment and generates faster time to insights for making better decisions. As a result, organizations can now quickly and efficiently delve deep into Hadoop to make decisions based on fact based analytical insights into all of the data for competitive advantage.

This session will focus on the data management and data preparation steps needed to fuel the analytical capabilities of your SAS on Hadoop solution, whether it is SAS Visual Analytics, SAS Visual Statistics or SAS In-Memory Statistics for Hadoop. You will learn which solutions are available to manage data in Hadoop, prepare data for In-Memory analysis and the various options that are for managing data quality, model scoring and executing SAS code, all running natively in a Hadoop cluster.