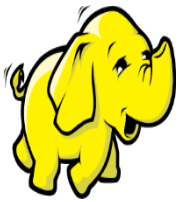
The Elephant in the Room: Hadoop & SAS Integration

Greenplum, SAS & Hadoop

Chris Stephens
Global Director of Analytics
Greenplum

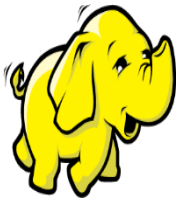


GREENPLUM®

EMC²

Agenda

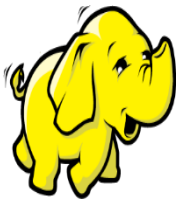
- Why is a Greenplum guy here talking about SAS & Hadoop?
- What is Hadoop anyway?
- Hadoop (huh!) what is it good for?
- Oh, and who is Greenplum?
- How do Greenplum and SAS work (together) with Hadoop



Introduction

- Chris Stephens
  - Global Director of Analytics, EMC/Greenplum

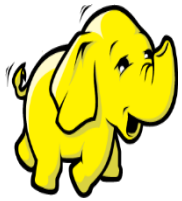
- 5 years experience working at SAS
  - Product Manager for High Performance Analytics
  - Product Manager for Model Manager
  - Technical Architect
- 12 years prior as SAS practitioner

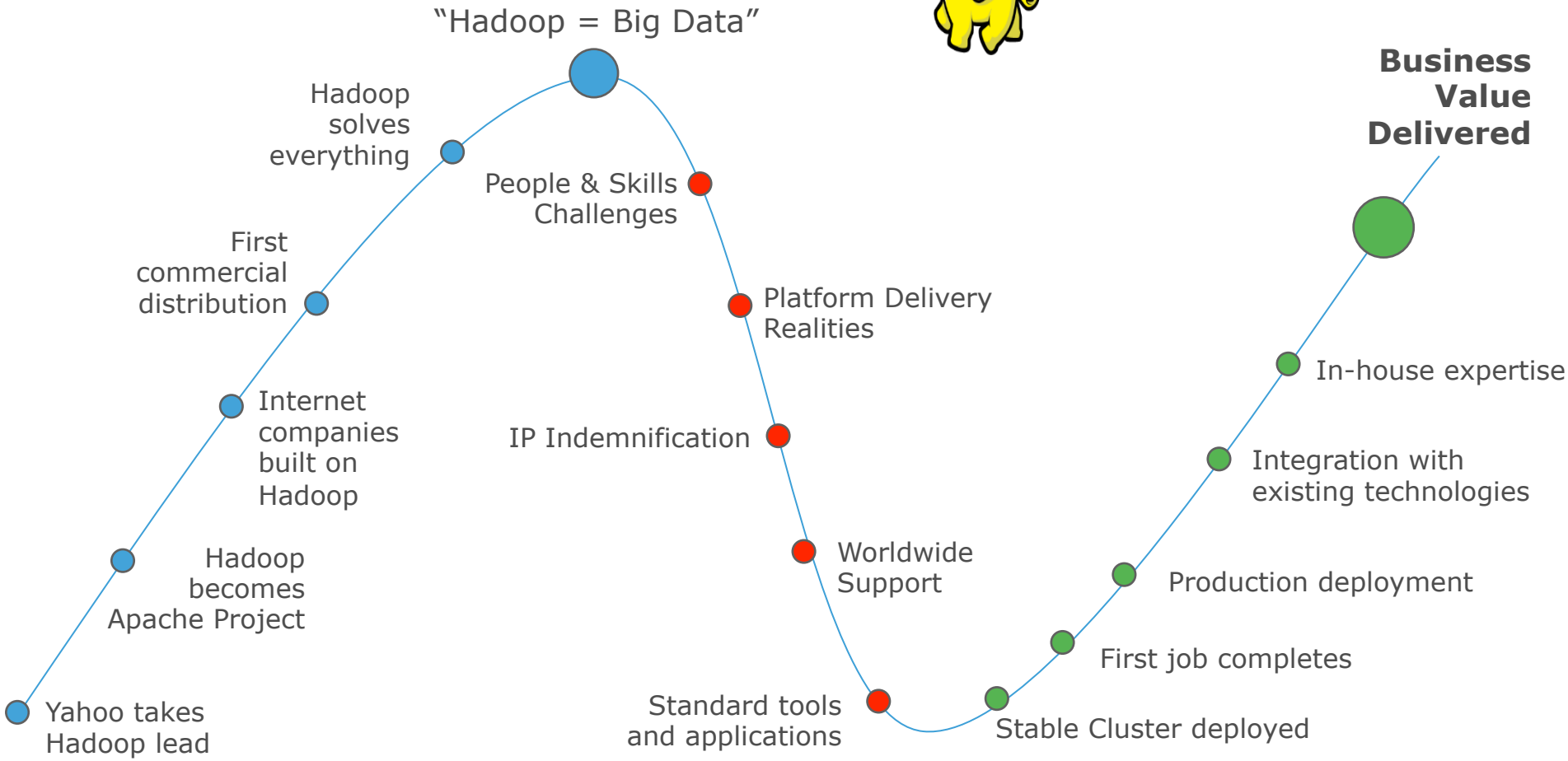


What is Hadoop?

- It's a framework for large-scale data processing:
  - Inspired by Google's architecture: MapReduce and GFS
  - A top-level Apache project–
    - Hadoop is open source
  - Written in Java, plus a few shell scripts
- Now the Hadoop "platform" consists of a tools set (Pig, Hive, etc)

The Hadoop Hype Cycle





What is it Good For?

- When you must process lots of unstructured data
- When your processing can easily be made parallel
- When running batch jobs is acceptable
- When you have access to lots of cheap hardware

- Indexing log files
- Processing Web Streams
- Image Files

What is it **Not So** Good For?

- For intense calculations with little data
- When your processing cannot be easily made parallel
- When your data is not self-contained
- When you need interactive results

- Relational Database Replacement
- Figuring Pi to 1,000,000 digits
- Analyzing numerical datasets

Greenplum Becomes the Foundation of EMC's Big Data Analytics (July 2010)

EMC ACQUIRES GREENPLUM

"For three years, Gartner has identified Greenplum as **the most advanced vendor** in the Visionary quadrant of its data warehouse DBMS Magic Quadrant…."

– Gartner

Greenplum, A Division of EMC

- 10 years of experience building and supporting enterprise-class massively parallel data processing software based on open source technology
- Silicon-valley based core engineering talent from Yahoo!, Teradata, Vertica, Netezza, Oracle, Amazon, Microsoft, IBM, and others
- 700 (and growing) personnel focused on Greenplum's Big Data Platform
  - Greenplum Database
  - Greenplum HD (Hadoop)
  - Chorus
  - Data Computing Appliances
  - Data Scientists
  - Pivotal Labs

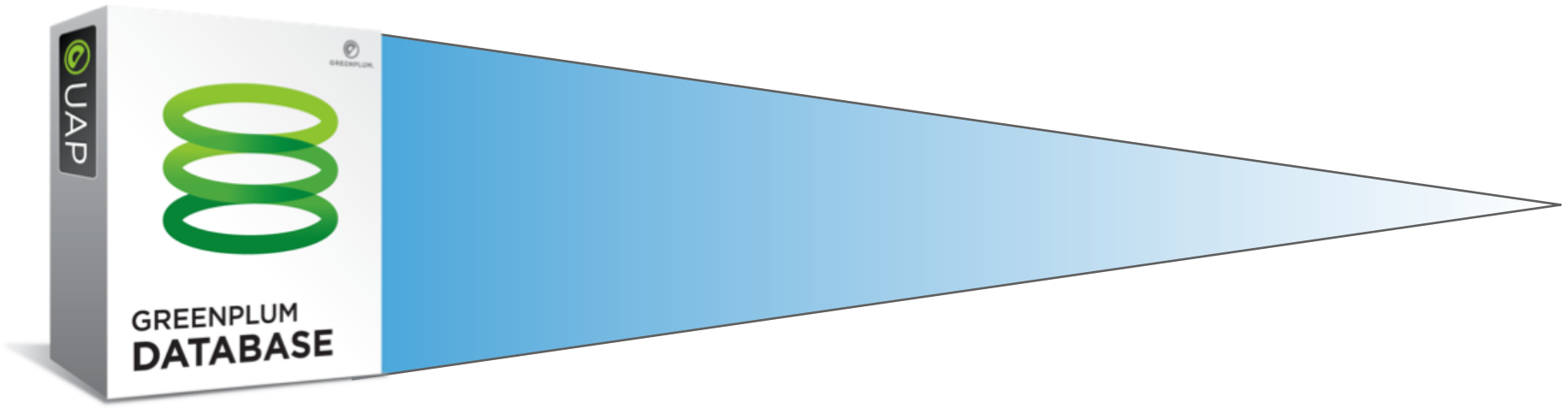
Greenplum Database

STRUCTURED **UNSTRUCTURED**

SQL Partitioning BI Tools

RDBMS Indexing Greenplum MapReduce

Tables and Schemas



Greenplum Hadoop

STRUCTURED **UNSTRUCTURED**

Schema on load

SequenceFile

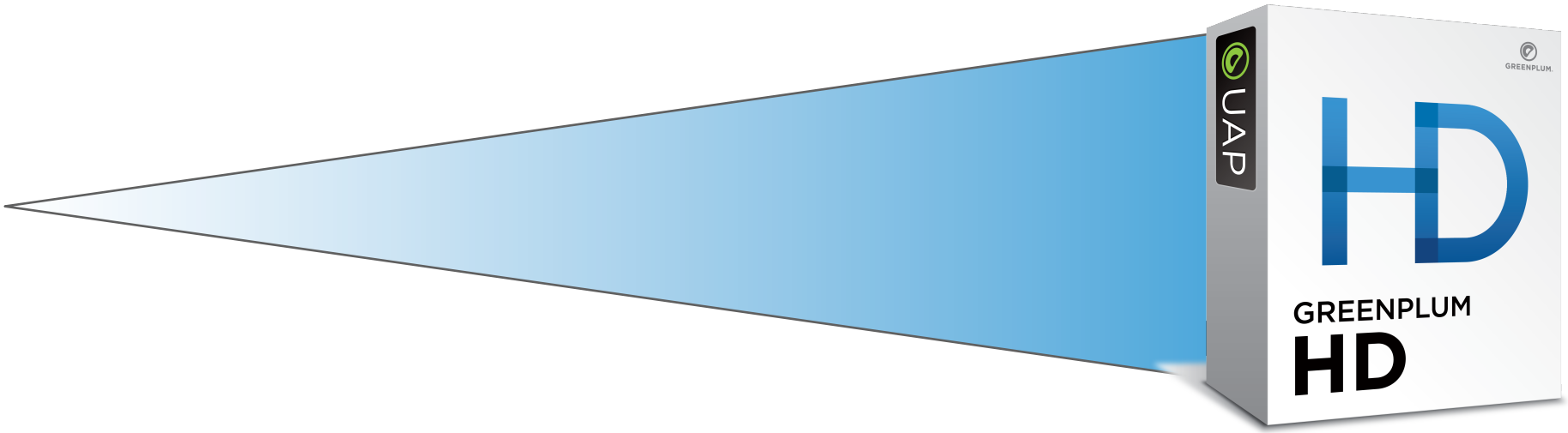
Directories MapReduce

Hive

Java

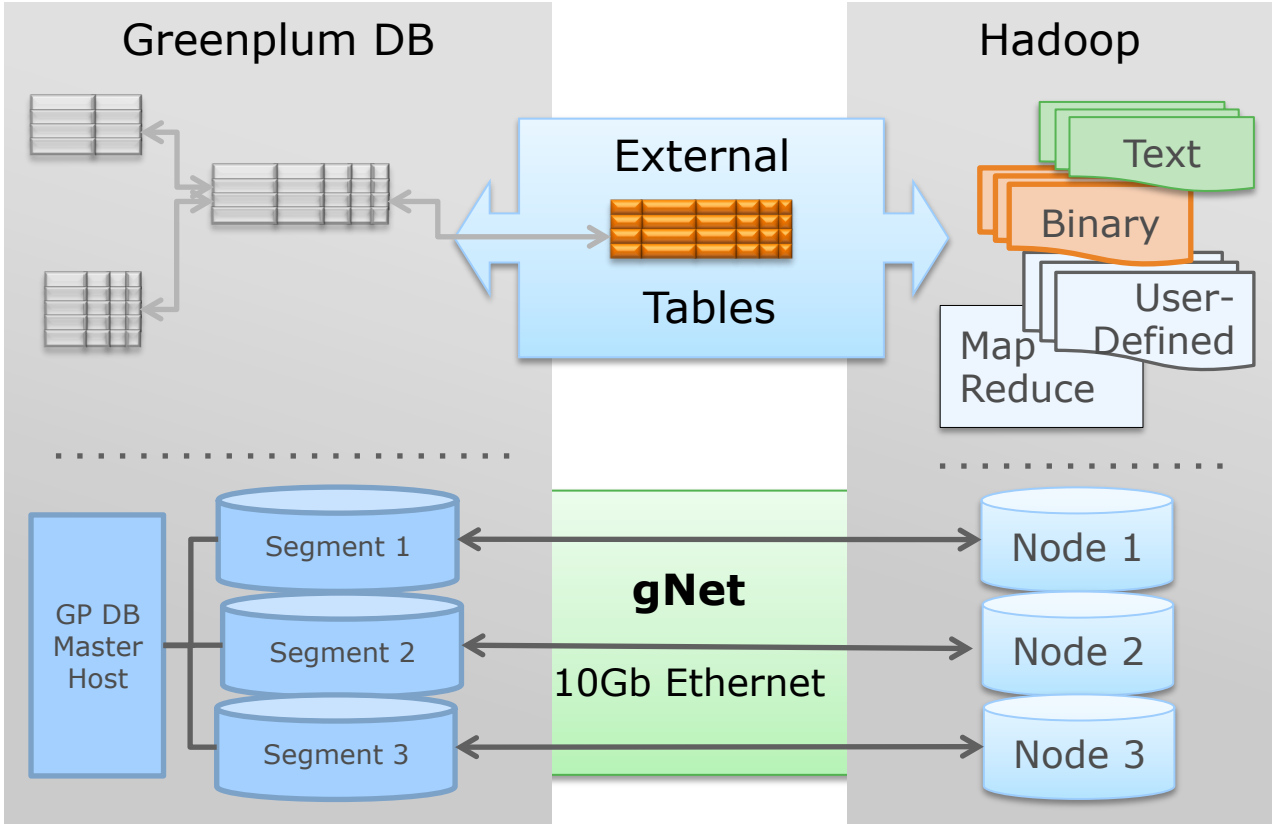
XML, JSON, … No ETL

Pig Flat files



Co-Pressing: gNet for UAP

Massively Parallel Access and Movement

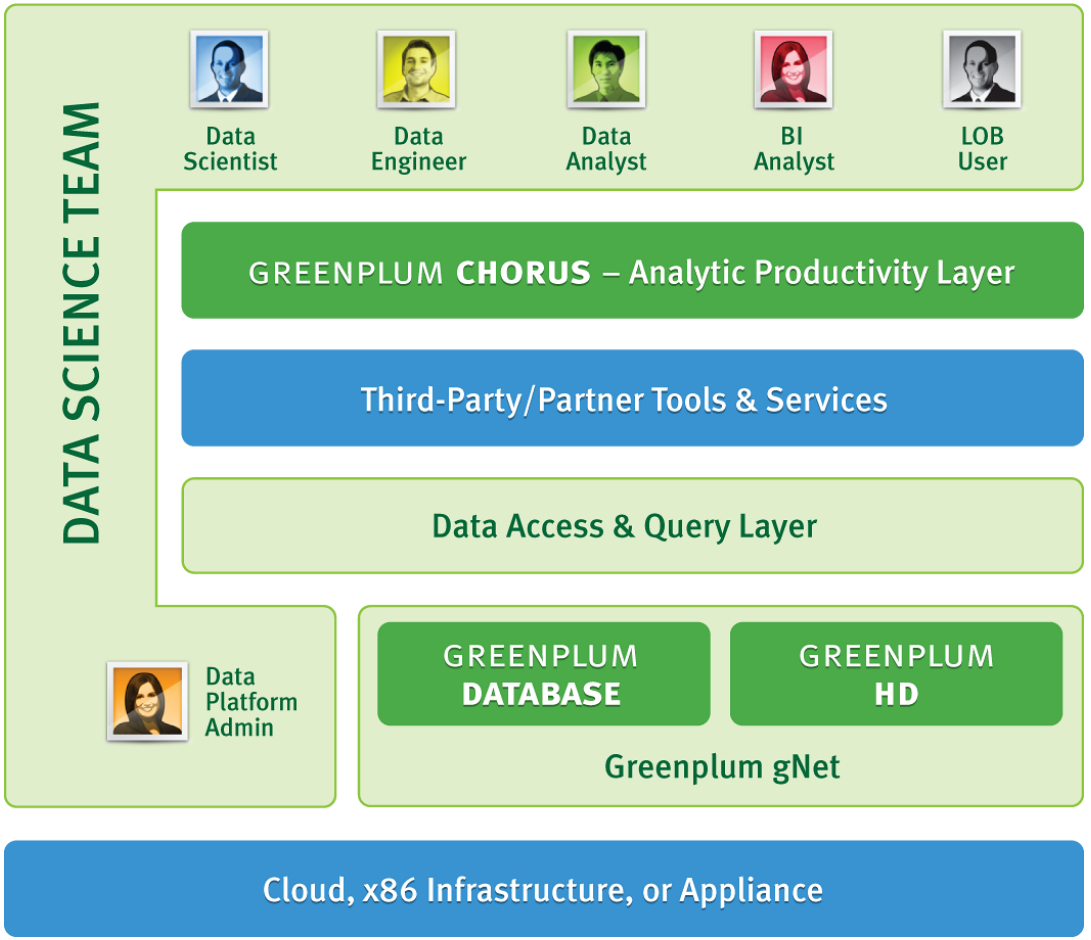


Custom formats via user-written MapReduce Java program And GPDB Format classes

- Maximize Solution Flexibility
- Minimize Data Duplication
- Access Hadoop Data in Real Time From Greenplum DB
- Import and export in Text, Binary and Compressed Formats

Delivered in a Unified Platform

- One system for Multi-structured analysis
- MPP Performance for data load and query
- Massive Scale
- Unified Collaboration, Management & Monitoring



Greenplum Investment in Hadoop Resources

- Total investment of over 100 EMC Full Time Personnel and Growing
  - Worldwide 24 X 7 customer support
  - Hadoop Research & Development team
  - Product Management staff from core Yahoo Engineering and Yahoo Labs teams responsible for developing and operationalizing the Yahoo's Hadoop clusters
  - Global field architects certified on Hadoop
  - Customer focused Emerging Technology Organization

Addressing the Talent Gap

- Hadoop Architecture Services
  - Installation and best practices
  - Educate the team
- Greenplum Analytics Labs
  - Leverage the expertise of Greenplum's Data Scientists
  - Packaged solutions that produce business value and actionable results
  - Accelerate Hadoop capabilities on your data with your analysts
- Establish a strategic vision
  - Roadmap for Hadoop and unified analytics

SAS + Greenplum: 5 Innovations

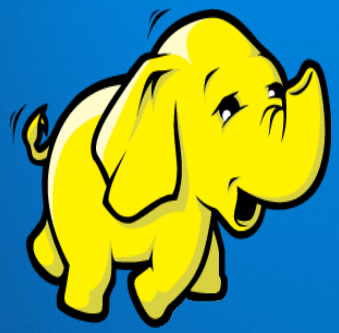
A Strategic Partnership for High-Performance Computing

- SAS/Access for Greenplum
  - Fast, transparent and secure access to Greenplum data from SAS
- SAS/Access for Hadoop
- SAS High-Performance Analytics for Greenplum
  - Runs SAS Analytics at In-Memory Speeds
  - Record-breaking scalability and performance.
- SAS Scoring Accelerator for Greenplum
  - Execute SAS Models in Parallel In-Database.
- SAS Grid for Greenplum
  - Dedicate Complete Servers
  - Integrate Into Modular Appliance
  - Connect via Appliance's Fast Backplane

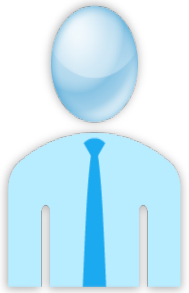
SAS & Hadoop

KEY TENANTS OF STRATEGY

1. CONTINUITY OF BUSINESS FOR SAS CUSTOMERS
2. COMPLIMENT HADOOP BY FILLING GAPS
3. LEVERAGE COMPONENTS FOR NEW SAS TECHNOLOGY



SAS® WITH HADOOP ECOSYSTEM



SAS® User

Next-Generation SAS® User

User Interface

SAS® Display Manager | SAS® Enterprise Guide® | SAS® Data Integration | SAS® Enterprise Miner™ | SAS® Visual Analytics

Metadata

SAS Metadata

Data Access

Base SAS & SAS/ACCESS® Interface to Hadoop™

In-Memory Data Access

Data Processing

HBase | Pig | Hive

MapReduce

sas

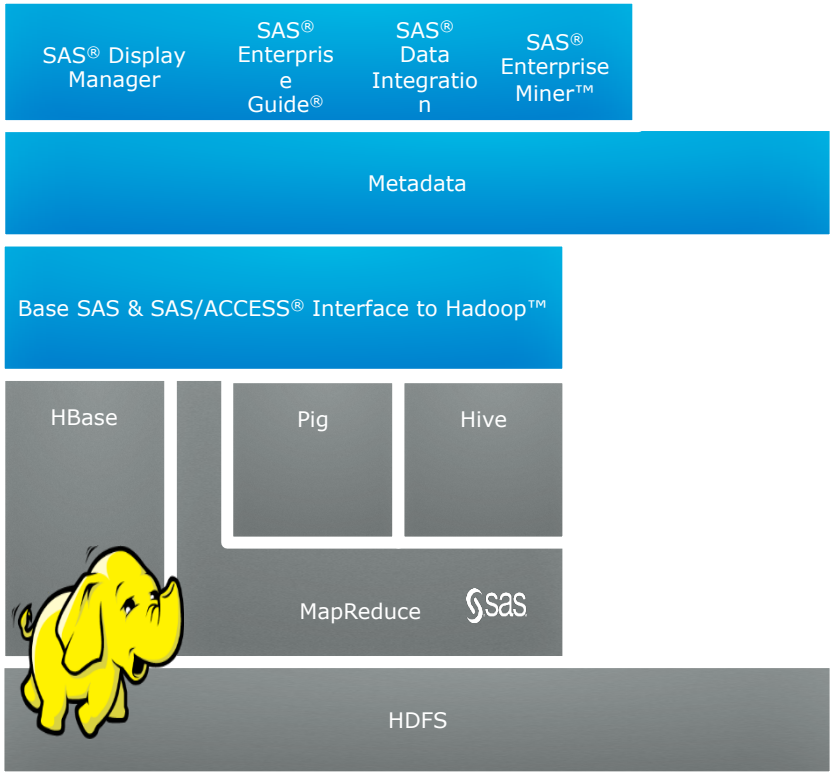
SAS® LASR™ Analytic Server & SAS® High-Performance Analytics

MPI Based

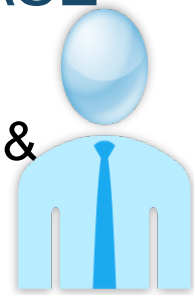
File System

HDFS

CONTINUITY OF BUSINESS

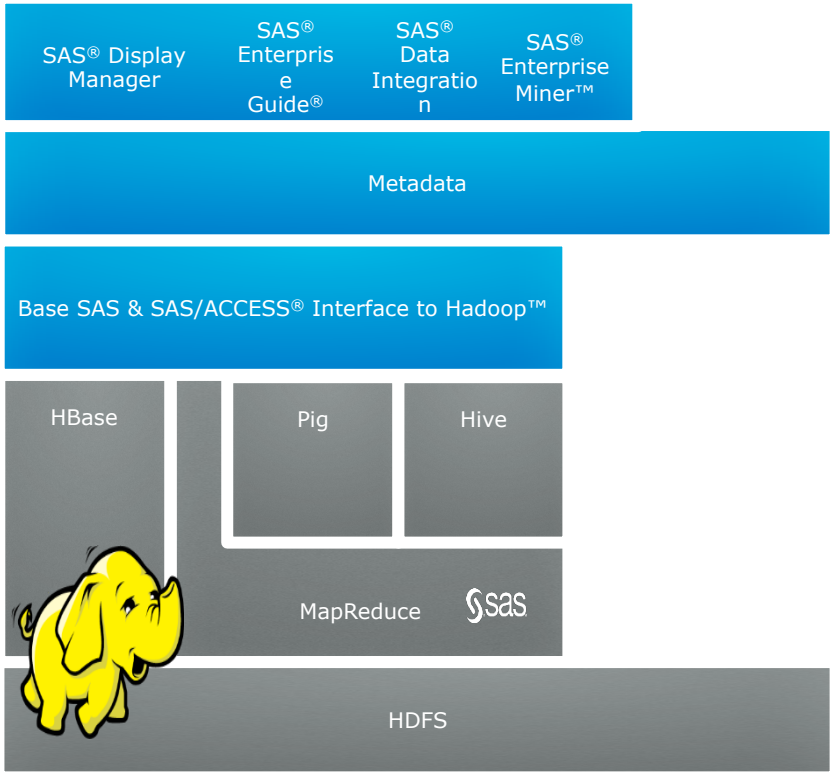


- SAS ACCESS TO HADOOP
  - Leverage Existing Investment in SAS Technologies
  - Minimize Training for End-Users to access data in Hadoop
- PROC HADOOP (BASE SAS)
  - Mix and Match SAS & Hadoop Processing

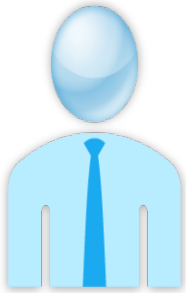


SAS® User

COMPLIMENT HADOOP



- DATA MANAGEMENT
  - SQL like Transformations for Hive
  - Pig, Hive and MR Transformations
- METADATA
  - Security, Lineage, Governance
- "HADOOP ACCELERATOR"s
  - Scoring Accelerator
  - Data Quality
  - Text Analytics



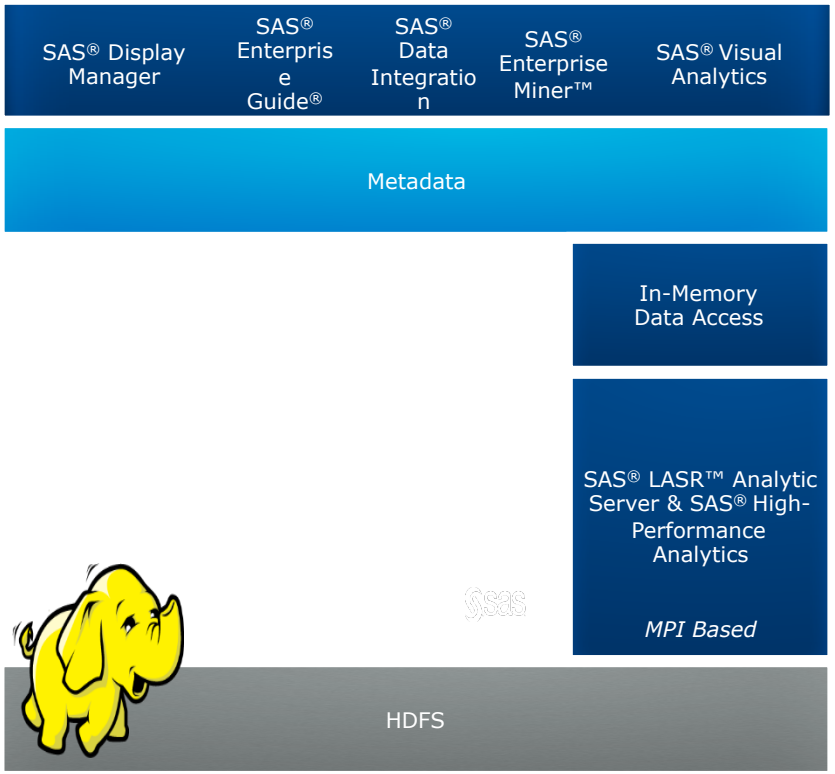
SAS® User

GREENPLUM®

EMC²

LEVERAGE HADOOP FOR NEW OFFERINGS

- SAS "Big Data" VISUAL ANALYTICS
  - Explorations
  - Visualizations, dashboards
  - Mobile Support
- SAS HIGH-PERFORMANCE ANALYTICS
  - "Big Data" Analytics Environment



Next-Generation SAS® User

GREENPLUM®

EMC²

BiG Data
Big Analytics

SAS HIGH PERFORMANCE ANALYTICS

Prepare

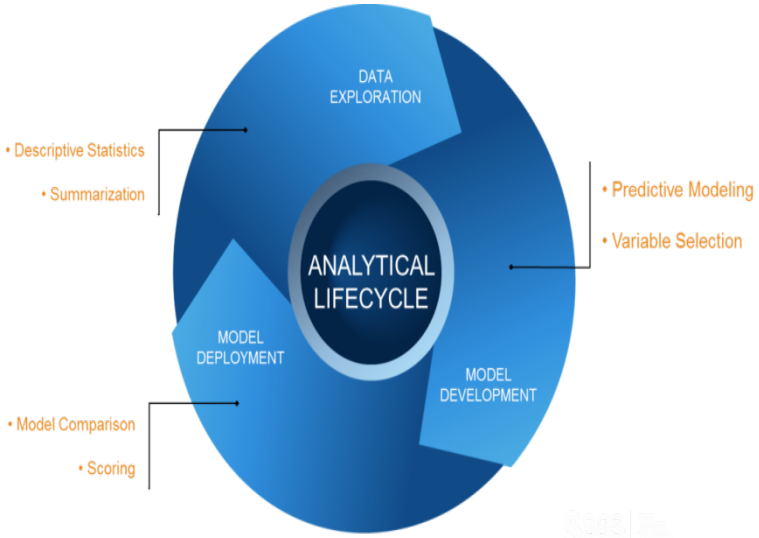
- HPDS2
- HPDMDB
- HPSAMPLE

Explore / Transform

- HPSUMMARY
- HPCORR
- HPREDUCE
- HPIMPUTE
- HPBIN

Model

- HPLOGISTIC
- HPREG
- HPNEURAL
- HPNLIN
- HPCOUNTREG
- HPMIXED
- HPSEVERITY
- HPFOREST
- HPSVM
- HPDECIDE
- HPQLIM
- HPLSO
- HPTMINE*
- HPTMSCORE*



GREENPLUM

EMC²

Greenplum Unified Analytic Platform