ANALYTICS EXPERIENCE 2016

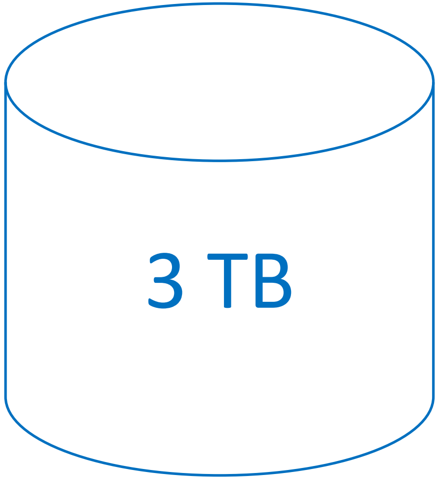
Getting Started with SAS and Hadoop

Jeff Bailey

#analyticsx

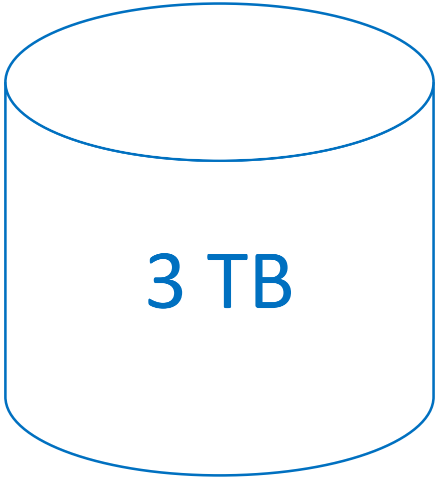
Why Hadoop?

HOW MUCH DOES THIS DRIVE COST?



HOW MUCH DOES THIS DRIVE COST?

Silly, you couldn’t get a 3TB drive in 1980!



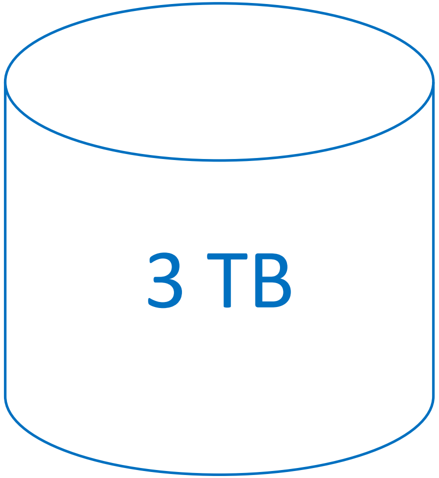
3 TB

1980

$1,312,500,000

HOW MUCH DOES THIS DRIVE COST?

That’s $0.03 per GB!



TODAY	$69
2010	$270
2005	$3,720
2000	$33,000
1995	$3,360,000
1990	$33,600,000
1985	$315,000,000
1980	$1,312,500,000

HOW MUCH DOES THIS DRIVE COST?

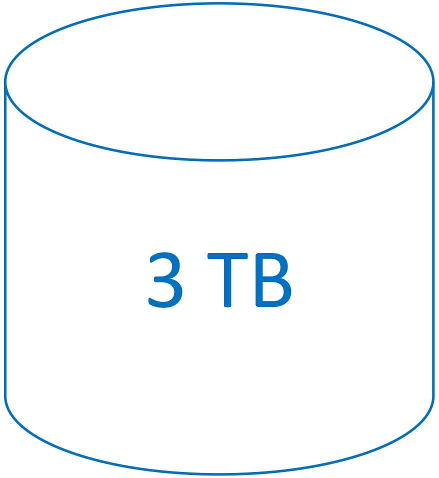
That’s $0.03 per GB!

TODAY	$69
2010	$270
2005	$3,720
1980	$1,312,500,000

Insight: Disk Space is FREE!

IT’S NOT JUST ABOUT COST!

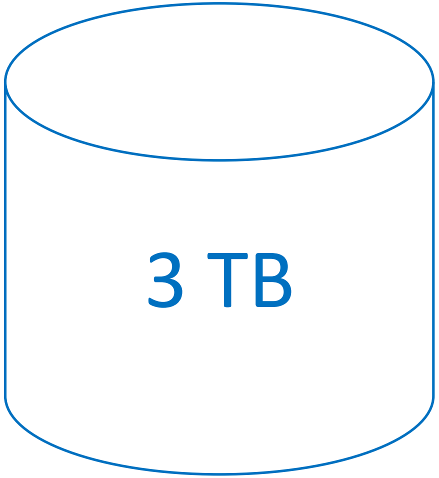
How long does it take to read **3 TB of data?**



3 TB

IT’S NOT JUST ABOUT COST!

How long does it take to read **3 TB of data?**



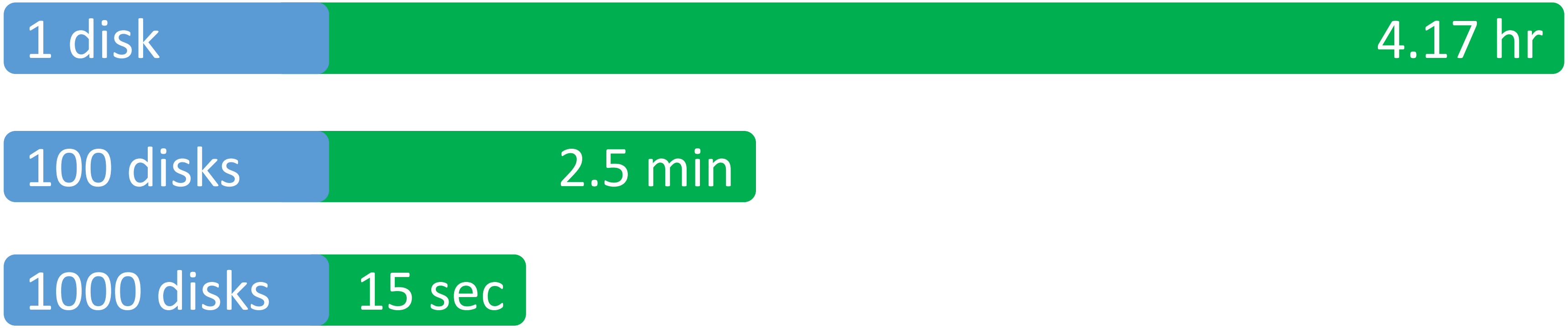
4.17 Hours

IT’S NOT JUST ABOUT COST!

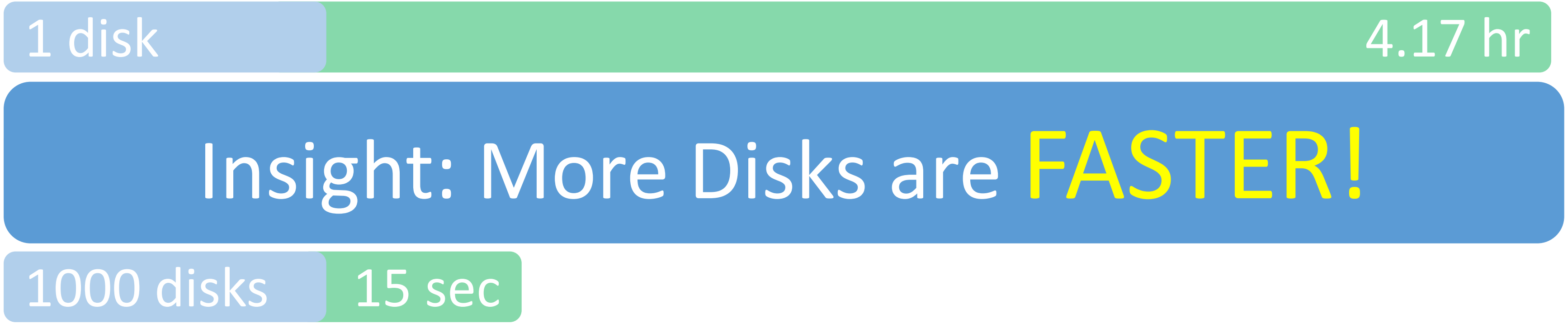
How long does it take [illegible] 2 TB?

What happens if you add more disks?

HOW LONG DOES IT TAKE TO READ A 3 TB FILE?



HOW LONG DOES IT TAKE TO READ A 3 TB FILE?



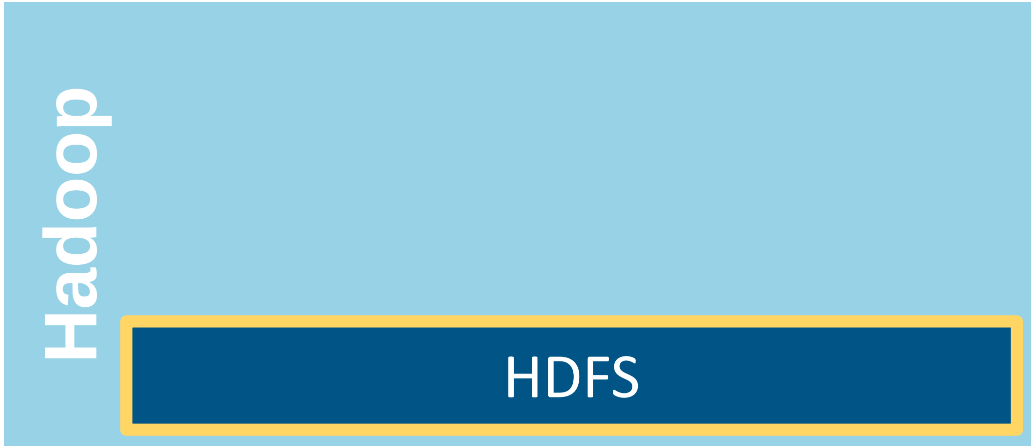
1 disk 4.17 hr

Insight: More Disks are FASTER!

1000 disks 15 sec

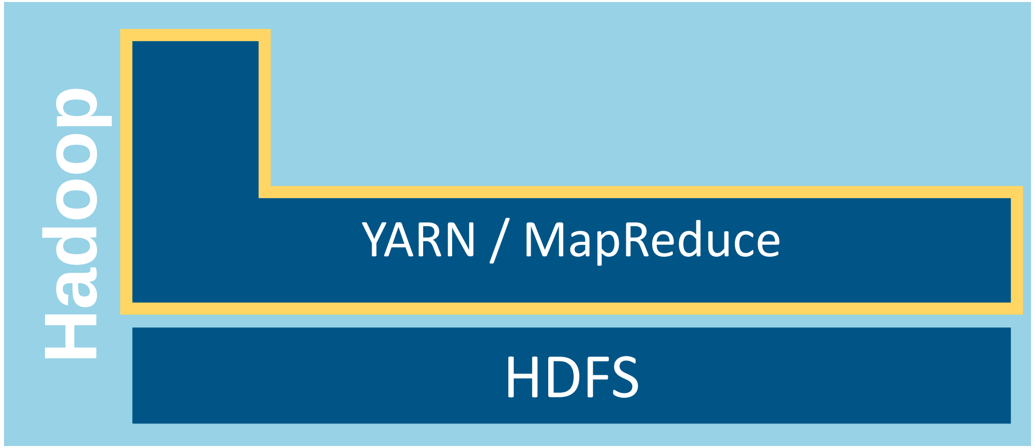
What is Hadoop?

Hadoop is a Storage Platform



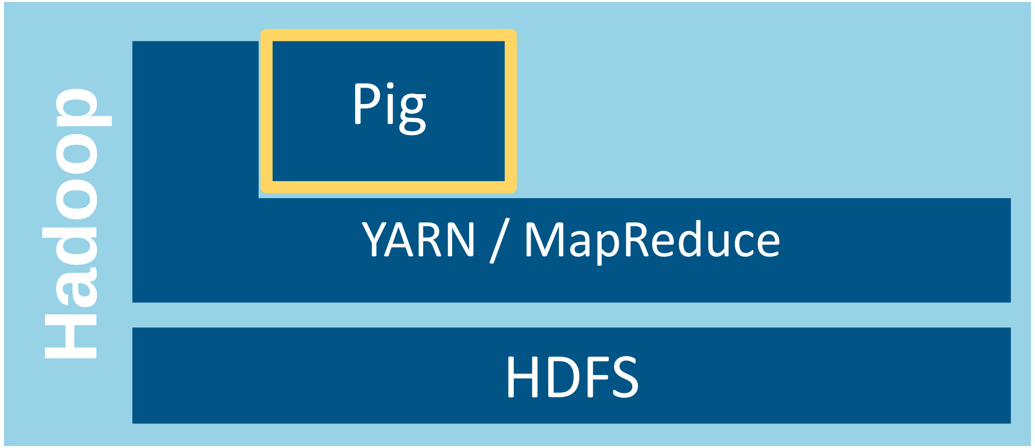
- Distributed Storage Performs Great
- Data is Replicated
- Reasonable Cost
- Sits on the OS File System

Hadoop is a Processing Platform



- MapReduce/YARN
- Distributed Processing
- Data Locality
- Usually Java

Apache Pig



- Scripting Language
- Higher level than programming Java MapReduce
- Pig Latin scripts are converted to MapReduce jobs
- Great for joining data
- Great for transforming data

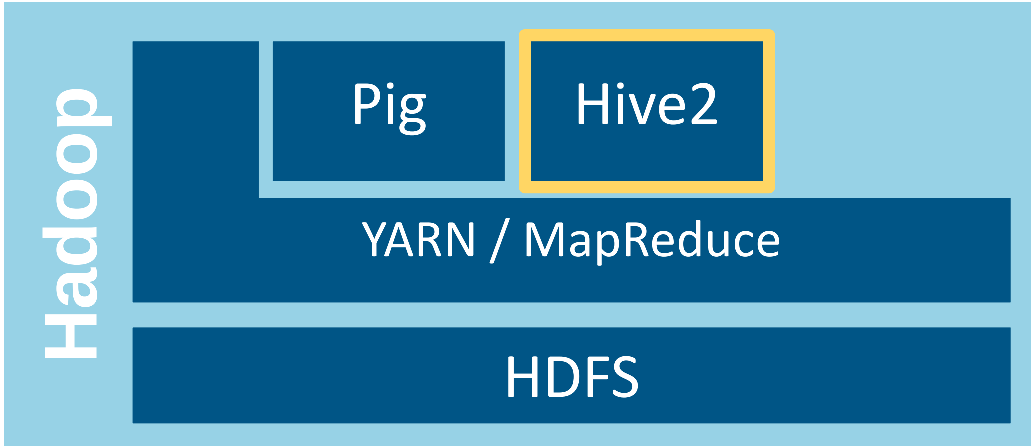
Apache Pig: Example Program

- Distributed Processing

Hadoop

```
people = LOAD '/user/training/customers' AS (cust_id, name);
orders = LOAD '/user/training/orders' AS (ord_id, cust_id, cost);
groups = GROUP orders BY cust_id;
totals = FOREACH groups GENERATE group, SUM(orders.cost) AS t;
result = JOIN totals BY group, people BY cust_id;
DUMP result;
```

Apache Hive



- SQL on Hadoop
- Similar to traditional SQL
- Reduces development time
- Enables BI on Hadoop
- Schema-on-Read
- You choose underlying file format

Apache Hive

- SQL on Hadoop

al

ent

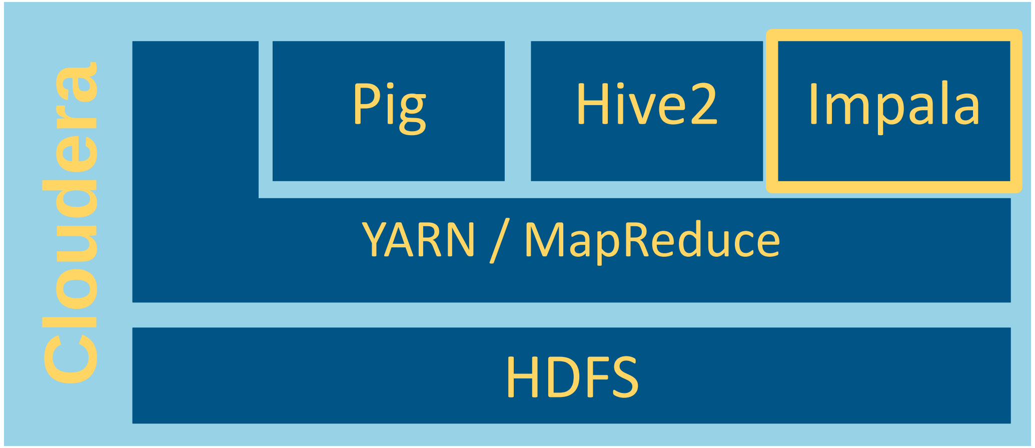
oop

ying

Hadoop

```
SELECT zipcode, SUM(cost) AS total
FROM customers
JOIN orders
ON (customers.cust_id = orders.cust_id)
WHERE zipcode LIKE '63%'
GROUP BY zipcode
ORDER BY total DESC;
```

Apache Impala is a SQL Engine

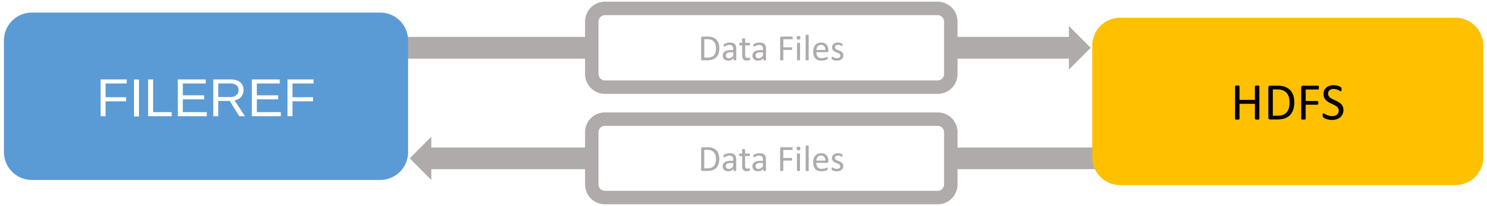


- High-performance SQL engine
- Handles concurrency well
- Does not rely on MapReduce
- Supports a dialect of SQL very similar to Hive’s
- 100% open source
- Apache License

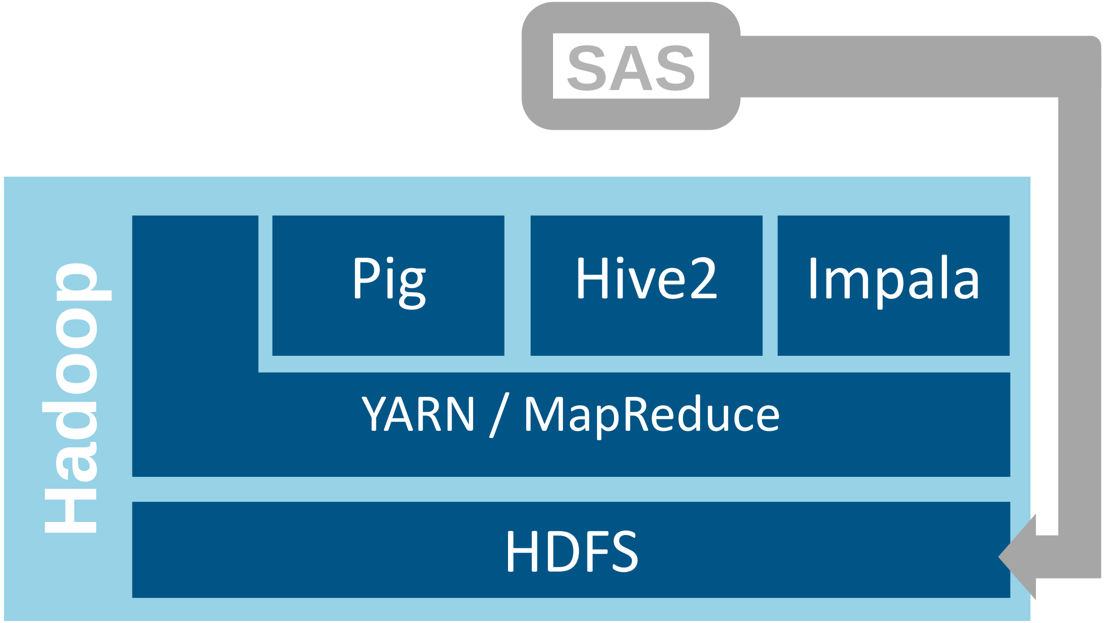
How can SAS Interact with Hadoop?

Using Base SAS 9.4 with Hadoop

#1



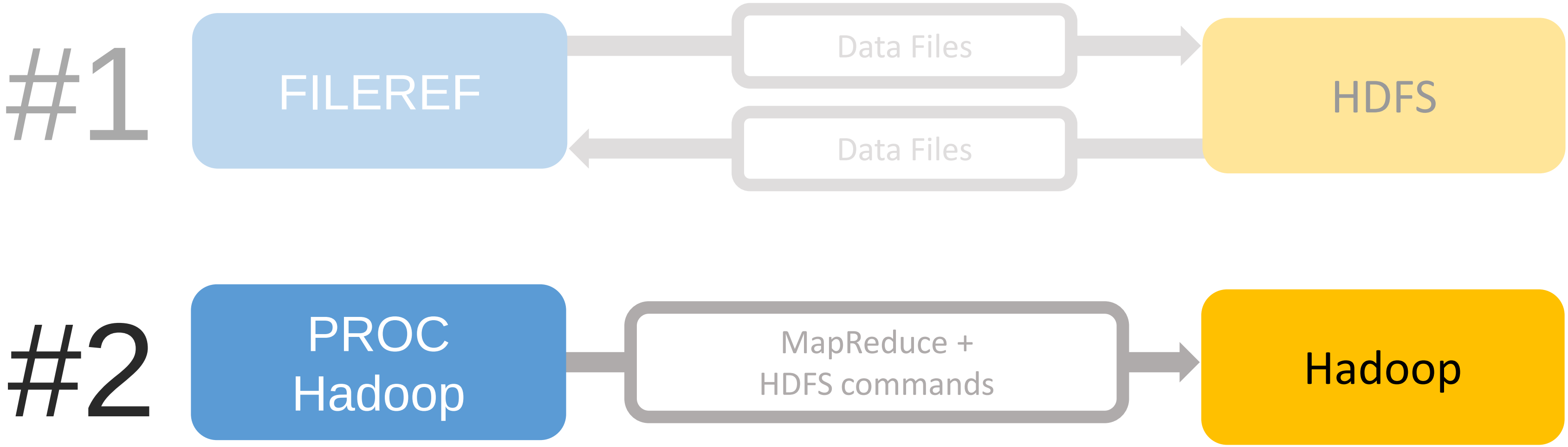
SAS FILENAME Statement for Hadoop



SAS FILENAME Statement for Hadoop

```
options set=SAS_HADOOP_CONFIG_PATH="\\sashq\cdh45p1";
options set=SAS_HADOOP_JAR_PATH="\\sashq\cdh45";

FILENAME hdp1 hadoop 'test.txt';

/* Write file to HDFS */
data _null_;
   file hdp1;
   put ' Test Test Test';
run;

/* Read file from HDFS */
data test;
   infile hdp1;
   input textline $15.;
run;
```

Using Base SAS 9.4 with Hadoop



#1

FILEREF

Data Files

Data Files

HDFS

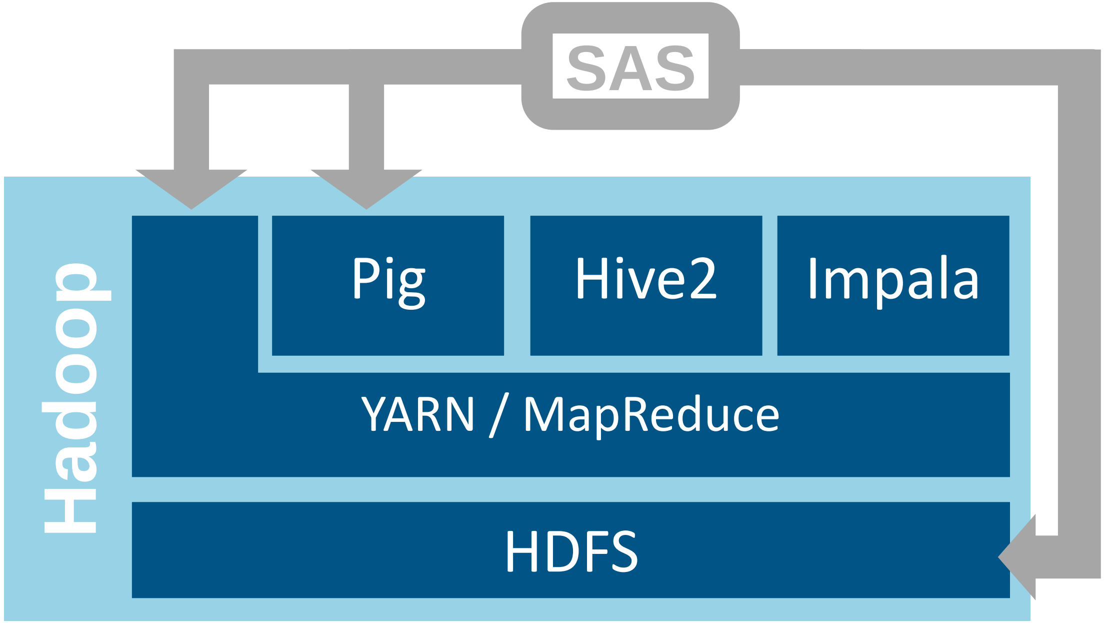
#2

PROC Hadoop

MapReduce + HDFS commands

Hadoop

Hadoop Procedure



- Submit HDFS commands
- Submit MapReduce Jobs
- Submit Pig Latin programs

How Do I Submit HDFS Commands?

```
filename cfg 'C:\Hadoop_cfg\cdh57.xml';

/* Copy war_and_peace.txt to HDFS. */
/* Copy moby_dick.txt     to HDFS. */
proc hadoop options=cfg username="sasxjb" verbose;
   HDFS mkdir='/user/sasxjb/Books';
   HDFS COPYFROMLOCAL="C:\Hadoop_data\moby_dick.txt"
             OUT='/user/sasxjb/Books/moby_dick.txt';
   HDFS COPYFROMLOCAL="C:\Hadoop_data\war_and_peace.txt"
             OUT='/user/sasxjb/Books/war_and_peace.txt';
run;
```

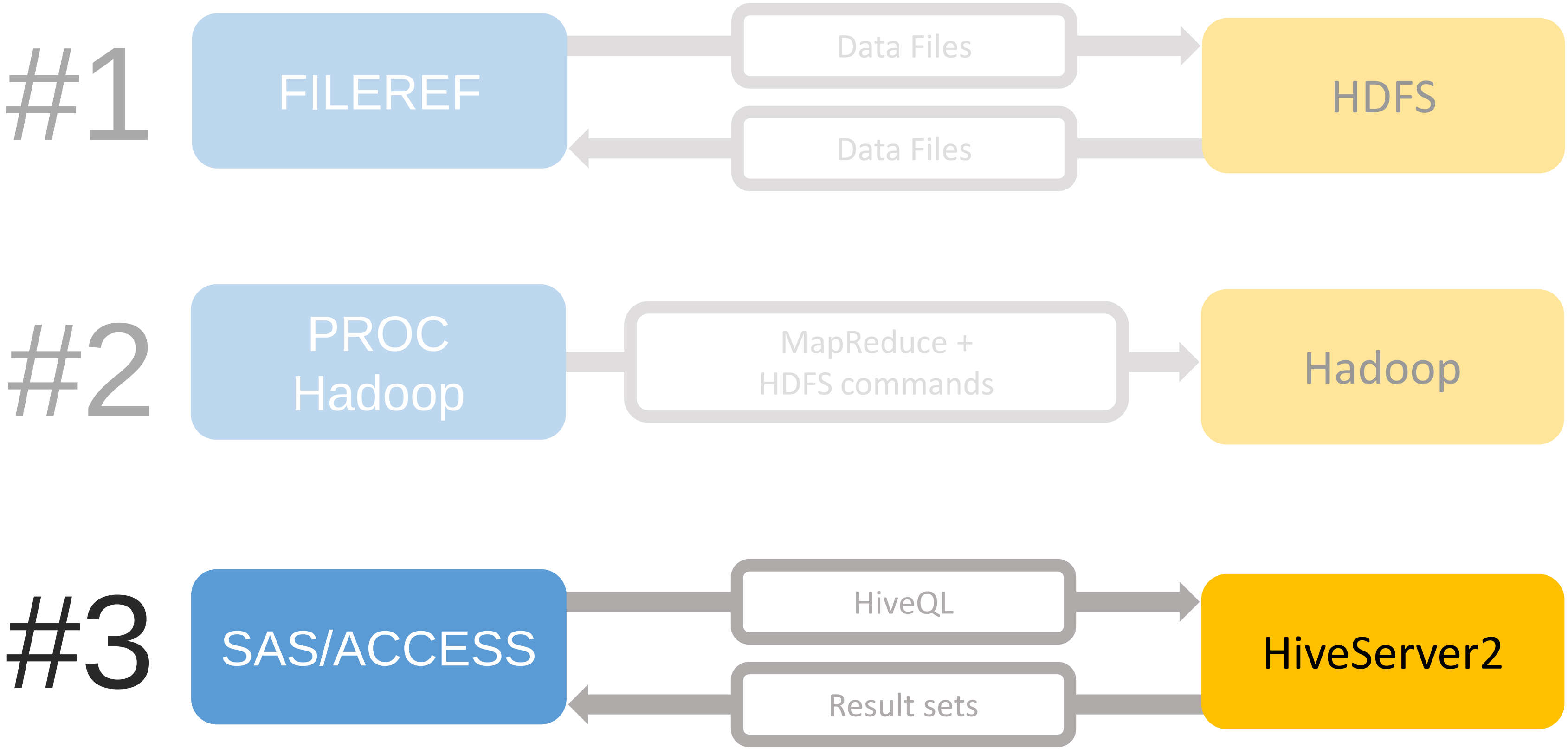
How Do I Submit MapReduce Jobs?

```
filename cfg 'C:\Hadoop_cfg\cdh57.xml';

proc hadoop options=cfg user="sasxjb" verbose;
  mapreduce input='/user/sasxjb/Books/moby_dick.txt'
    output='/user/sasxjb/outBook'
    jar='C:\Hadoop_examples\hadoop-examples-1.2.0.1.3-96.jar'
    outputkey="org.apache.hadoop.io.Text"
    outputvalue="org.apache.hadoop.io.IntWritable"
    reduce="org.apache.hadoop.examples.WordCount$IntSumReducer"
    combine="org.apache.hadoop.examples.WordCount$IntSumReducer"
    map="org.apache.hadoop.examples.WordCount$TokenizerMapper";
run;
```

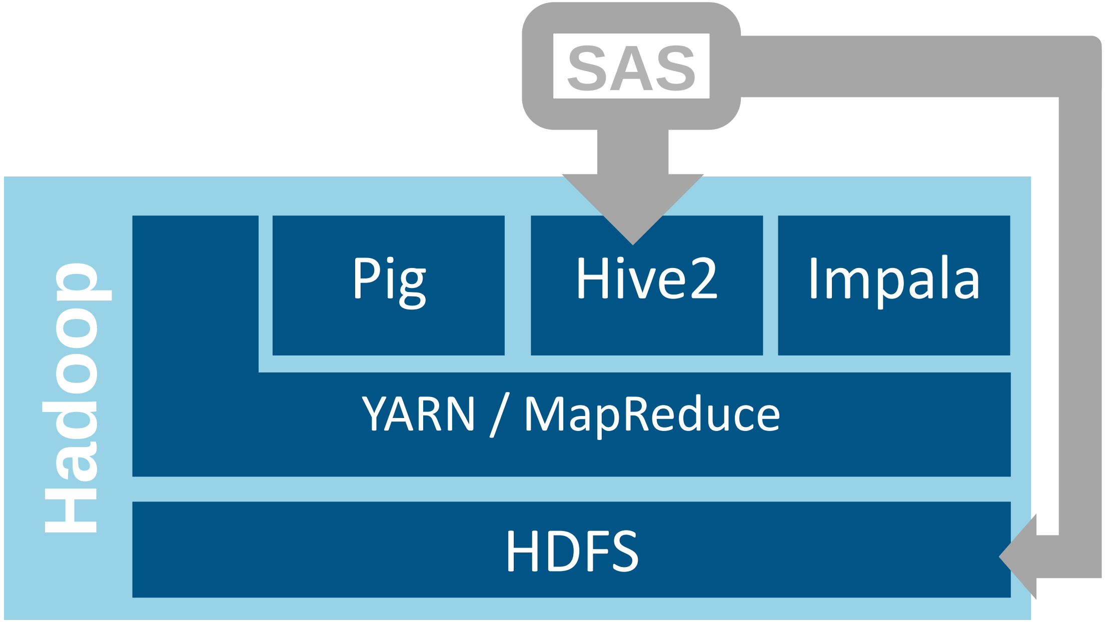
How Do I Submit Pig Latin Programs?

```
filename cfg 'C:\Hadoop_cfg\cdh57.xml';

proc hadoop options=cfg username="sasxjb“ verbose;
     pig code=pigcode ;
run;
```

Using Base SAS 9.4 with Hadoop



SAS/ACCESS Interface to Hadoop



- Connects via JDBC
- Makes Hive tables look like SAS data sets
- Bulk loads directly to HDFS
- Can read directly from HDFS

How Does SAS/ACCESS Talk to Hadoop?

```
proc sql;
    select count(*) from mycdh.customer_dim
        where loyalty_program='Chocolate Club';
run;
```

?

How Does SAS/ACCESS Talk to Hadoop?

```
proc sql;
   select count(*) from mycdh.customer_dim
      where loyalty_program='Chocolate Club';
run;
```

```
select COUNT(*) from `CUSTOMER_DIM` TXT_1
WHERE TXT_1.`loyalty_program` = 'Chocolate Club'
```

SAS Generated This SQL

How Does SAS/ACCESS Talk to Hadoop?

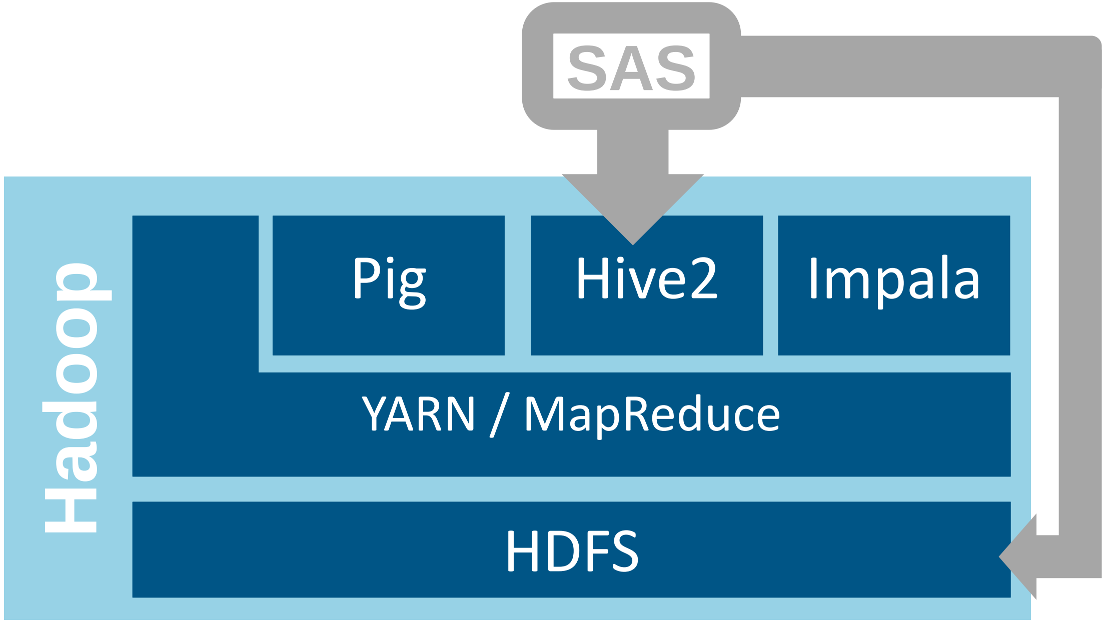
```
proc sql;
    select count(*) from mycdh.customer_dim
        where loyalty_program='Chocolate Club';
run;
```

```
OPTIONS  SASTRACE=',,,d' SASTRACELOC=SASLOG NOSTSUFFIX;
```

```
select COUNT(*) from `CUSTOMER_DIM` TXT_1
WHERE TXT_1.`loyalty_program` = 'Chocolate Club'
```

SAS Generated This SQL

SAS/ACCESS Interface to Hadoop



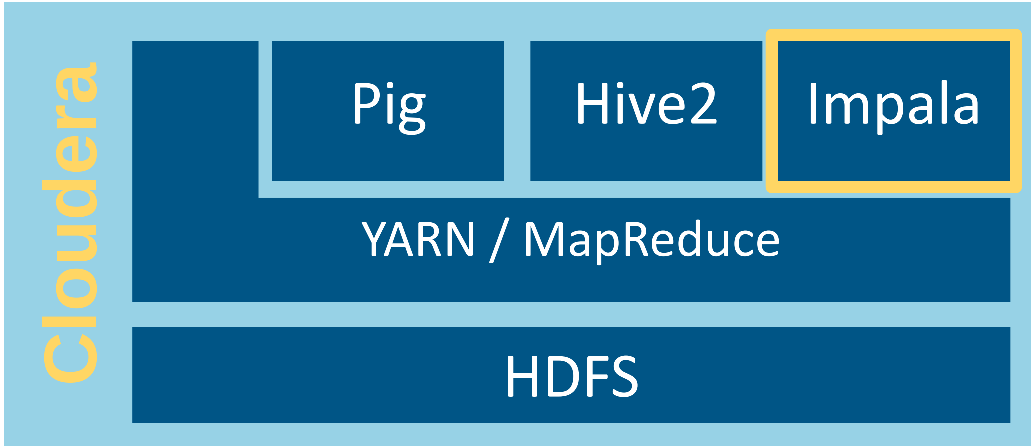
- **Generates HiveQL**
- Connects via JDBC
- Makes Hive tables look like SAS data sets
- Bulk loads directly to HDFS
- Can read directly from HDFS

We Can Write Our Own HiveQL!

```
proc sql;
   connect to hadoop (server=quickstart
                      user=cloudera);
   execute (create table store_cnt
                row format delimited
                fields terminated by '\001`
                stored as parquet
               as
                  select customer_rk, count(*) as tot
                    from order_fact
                   group by customer_rk) by hadoop;
quit;
```

Explicit Pass-Through

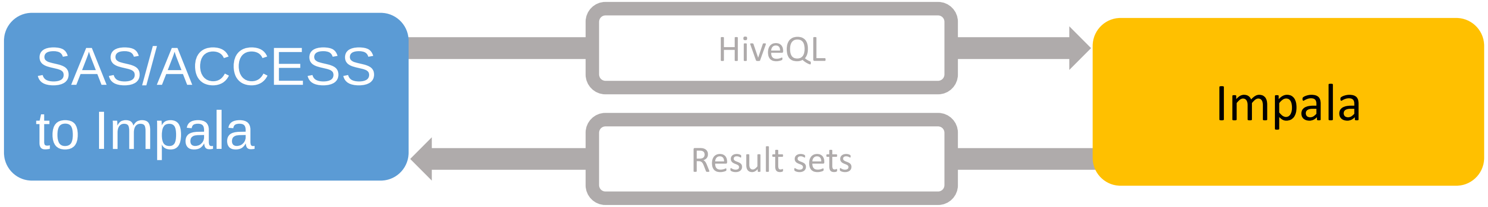
What about Apache Impala?



SAS/ACCESS Interface to Impala:

- Connects via ODBC
- Makes Hive tables look like SAS data sets
- Bulk loads directly to HDFS

#4



In-Database: Code Accelerator

What is SAS In-Database Code Accelerator?

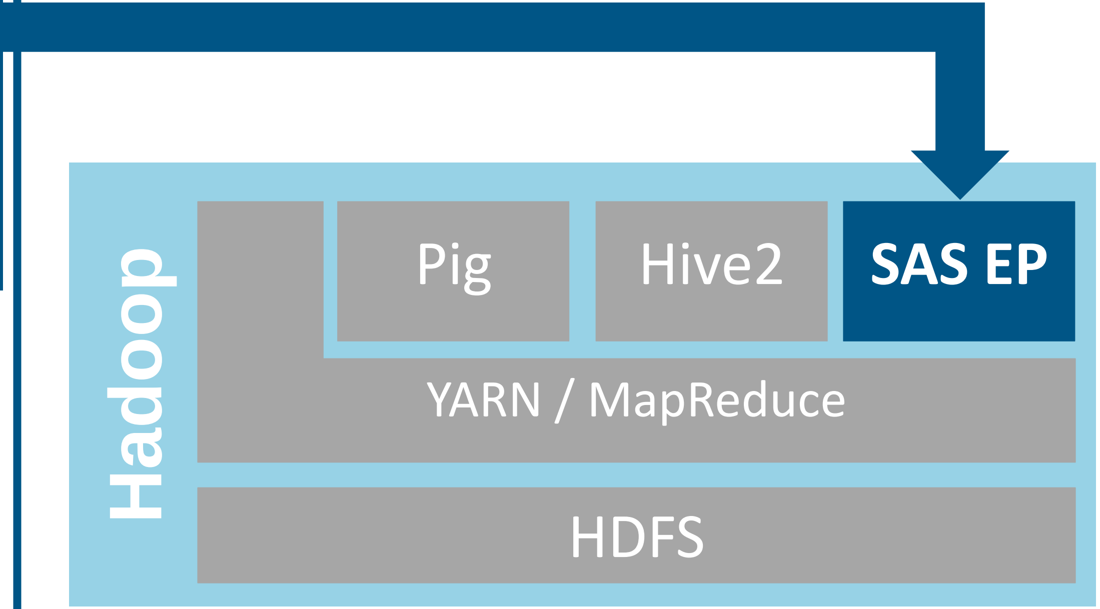
```
proc ds2 indb=yes;
   thread tpgm / overwrite=yes;
      method run();
        set hdplib.intable;
        output;
   end;
   endthread;
   run;
   data hdplib.outdata

(overwrite=yes);
      dcl thread tpgm hdpdata;
      method run();
         set from hdpdata;
      end;
   enddata;
   run;
quit;
```

SAS In-Database Code Accelerators let you run SAS code inside Hadoop. With this you get:

- DS2 processing (modern DATA Step)
- More Data Types
- Code Packages
- More Programming Structures
- Parallel Database Operations
- Thread Programs Run Inside Database

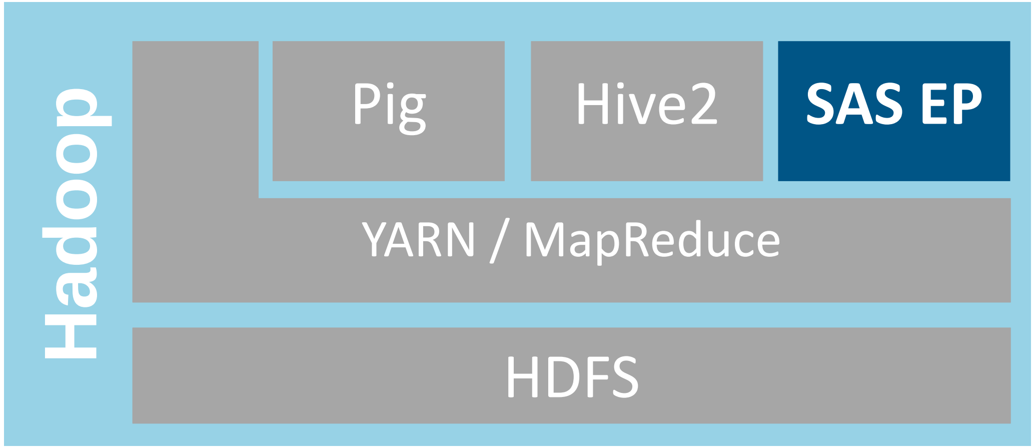
In-Database Code Accelerator Runs in Hadoop

```
proc ds2 indb=yes;
   thread tpgm / overwrite=yes;
      method run();
         set hdplib.intable;
         output;
   end;
   endthread;
   run;
   data hdplib.outdata

 (overwrite=yes);
      dcl thread tpgm hdpdata;
      method run();
         set from hdpdata;
      end;
   enddata;
   run;
quit;
```



In-Database: Scoring Accelerator

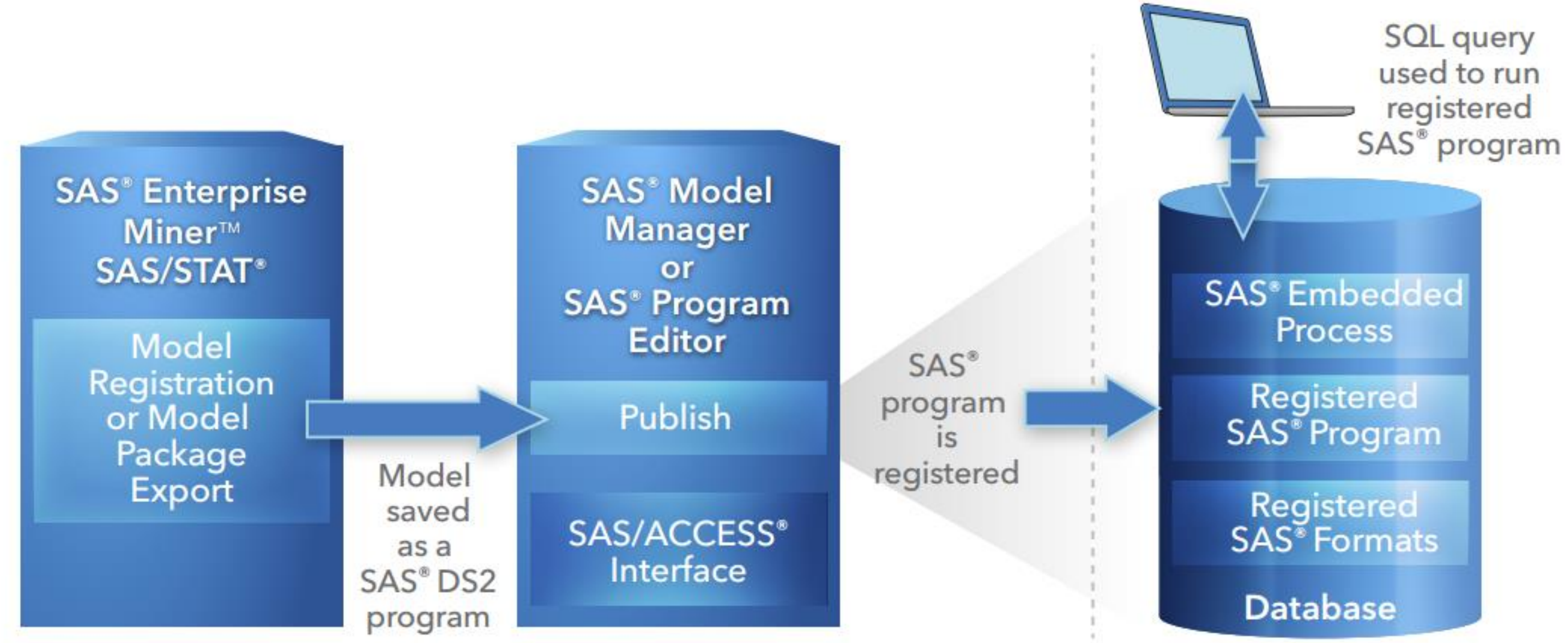
What is SAS In-Database Scoring Accelerator?



SAS In-Database Scoring Accelerator lets you score models inside the cluster. With this you get:

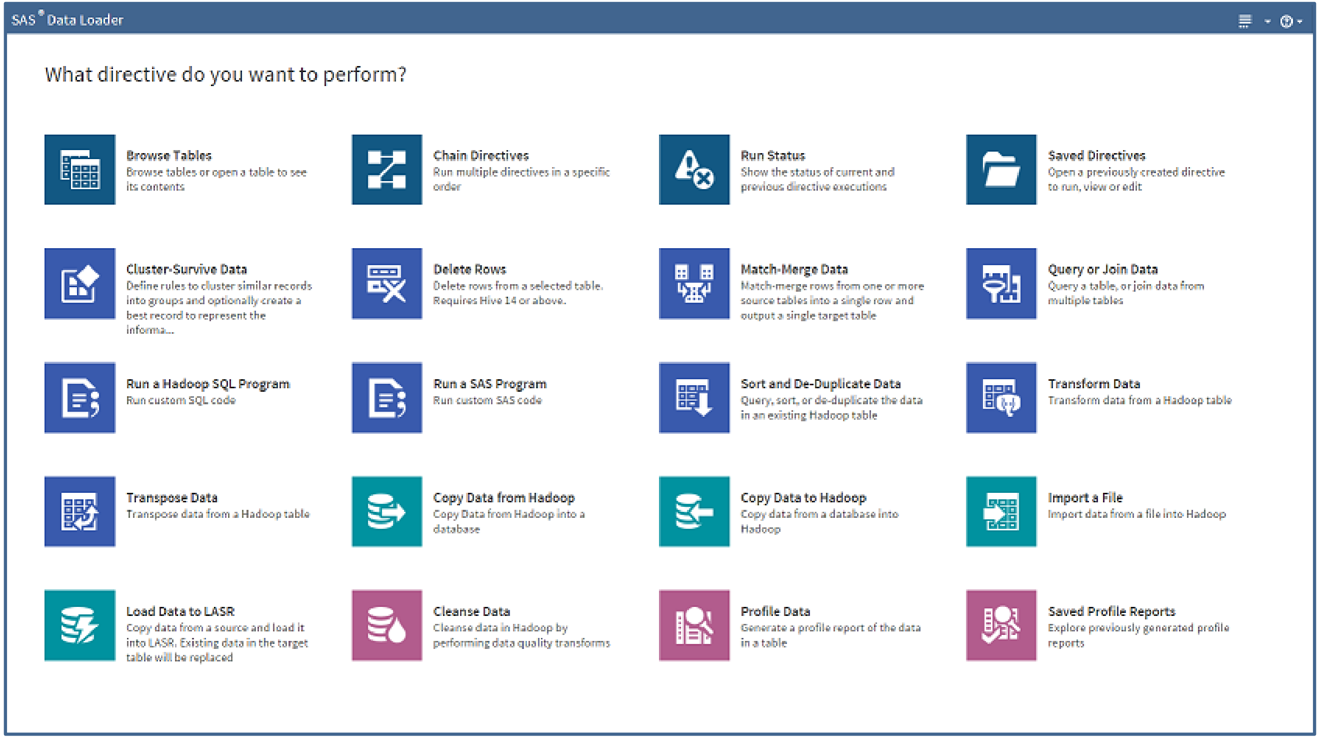
- Uses the SAS Embedded Process
- Faster Scoring
- Less data movement – score data where it lives
- Uses fewer resources

What does the Scoring Process Look like?


SAS® Enterprise Miner™ SAS/STAT®
Model Registration or Model Package Export
Model saved as a SAS® DS2 program
SAS® Model Manager or SAS® Program Editor
Publish
SAS/ACCESS® Interface
SAS® program is registered
SQL query used to run registered SAS® program
SAS® Embedded Process
Registered SAS® Program
Registered SAS® Formats
Database

Data Loader for Hadoop

Data Loader for Hadoop – Self Service Big Data



SAS® Data Loader

What directive do you want to perform?

Browse Tables
Browse tables or open a table to see its contents

Chain Directives
Run multiple directives in a specific order

Run Status
Show the status of current and previous directive executions

Saved Directives
Open a previously created directive to run, view or edit

Cluster-Survive Data
Define rules to cluster similar records into groups and optionally create a best record to represent the informa...

Delete Rows
Delete rows from a selected table. Requires Hive 14 or above.

Match-Merge Data
Match-merge rows from one or more source tables into a single row and output a single target table

Query or Join Data
Query a table, or join data from multiple tables

Run a Hadoop SQL Program
Run custom SQL code

Run a SAS Program
Run custom SAS code

Sort and De-Duplicate Data
Query, sort, or de-duplicate the data in an existing Hadoop table

Transform Data
Transform data from a Hadoop table

Transpose Data
Transpose data from a Hadoop table

Copy Data from Hadoop
Copy Data from Hadoop into a database

Copy Data to Hadoop
Copy data from a database into Hadoop

Import a File
Import data from a file into Hadoop

Load Data to LASR
Copy data from a source and load it into LASR. Existing data in the target table will be replaced

Cleanse Data
Cleanse data in Hadoop by performing data quality transforms

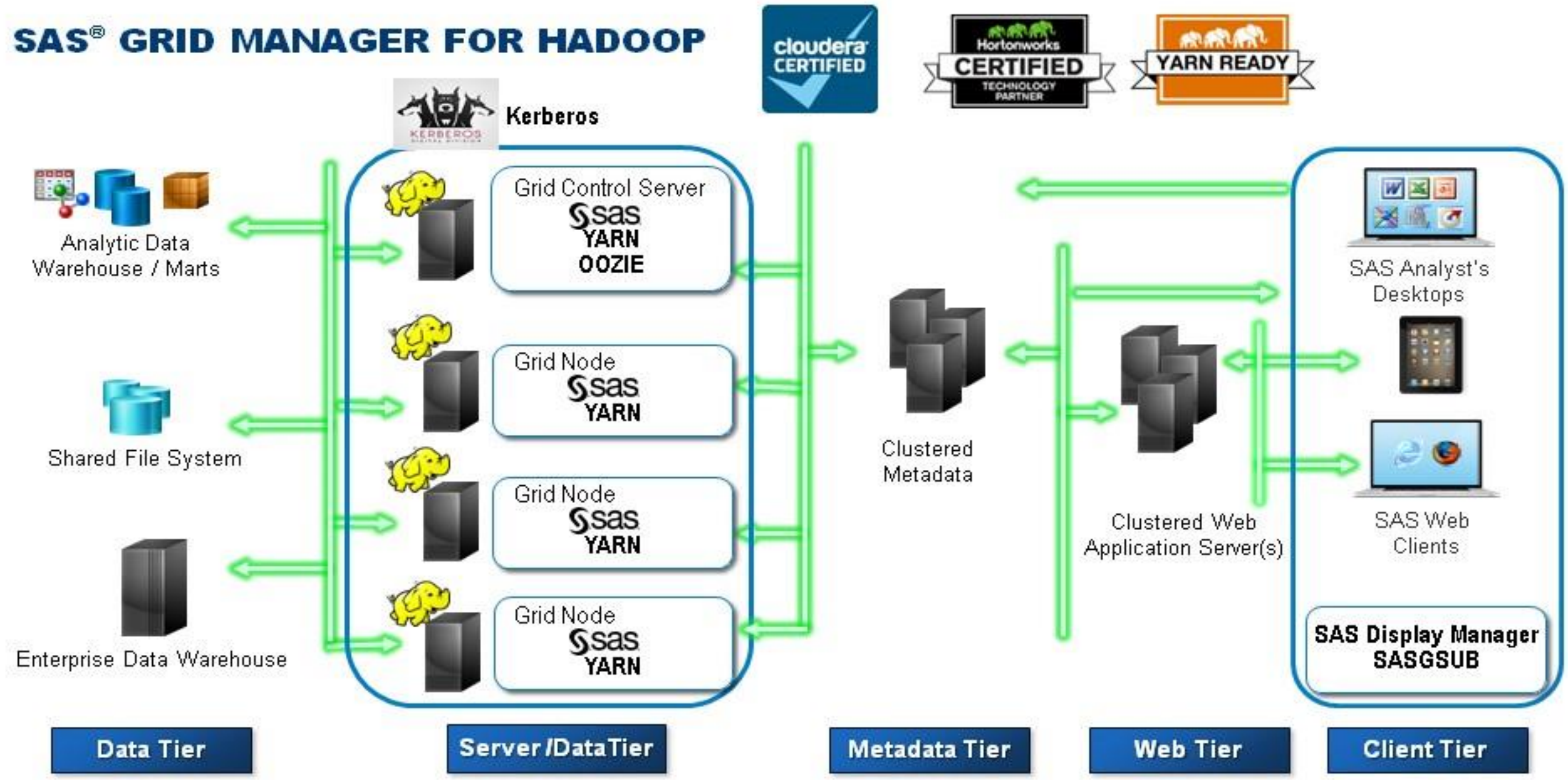
Profile Data
Generate a profile report of the data in a table

Saved Profile Reports
Explore previously generated profile reports

- Easy to use UI
- Query Data
- Manage Data
- Transform Data
- Run Custom Code
- Move Data

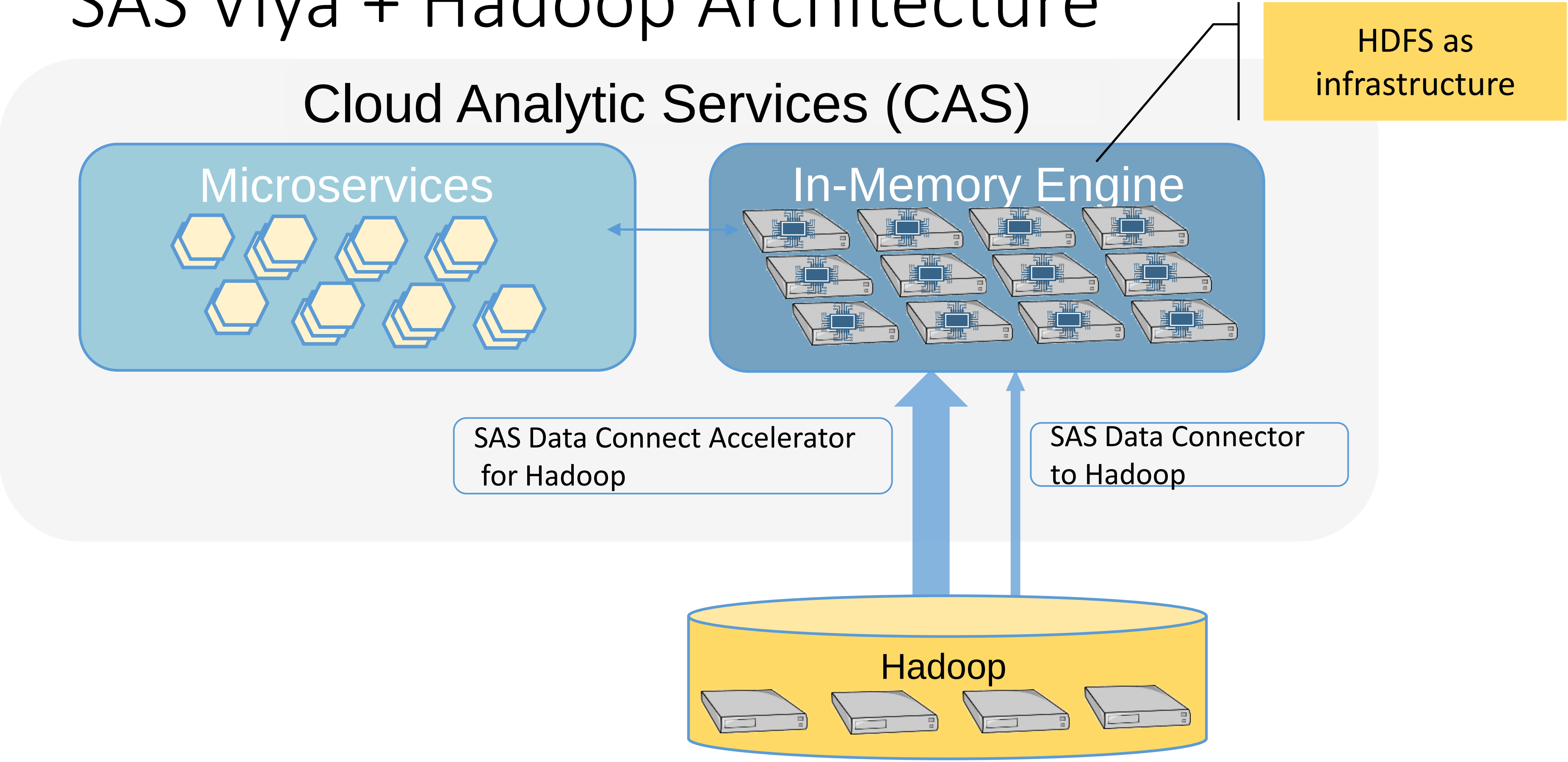
SAS Grid Manager for Hadoop

What is SAS Grid Manager for Hadoop?


SAS® GRID MANAGER FOR HADOOP
cloudera CERTIFIED
Hortonworks CERTIFIED TECHNOLOGY PARTNER
YARN READY
Kerberos
Analytic Data Warehouse / Marts
Shared File System
Enterprise Data Warehouse
Grid Control Server sas YARN OOZIE
Grid Node sas YARN
Grid Node sas YARN
Grid Node sas YARN
Clustered Metadata
Clustered Web Application Server(s)
SAS Analyst's Desktops
SAS Web Clients
SAS Display Manager SASGSUB
Data Tier
Server /DataTier
Metadata Tier
Web Tier
Client Tier

Servers

SAS Viya + Hadoop Architecture



HDFS as infrastructure
Cloud Analytic Services (CAS)
Microservices
In-Memory Engine
SAS Data Connect Accelerator for Hadoop
SAS Data Connector to Hadoop
Hadoop

Feel Free to Contact Me!

Jeff.Bailey@sas.com

http://www.linkedin.com/in/jeffreydbailey

https://github.com/Jeff-Bailey/SAS13341_SAS_Hadoop

ANALYTICS EXPERIENCE
2016

#analyticsx