JAMES COOK UNIVERSITY
AUSTRALIA

Week 5

MA5831 – Advanced Data Management and Analysis using SAS

Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au

online.jcu.edu.au

Cairns
Singapore
Townsville

JAMES COOK UNIVERSITY AUSTRALIA

Agenda

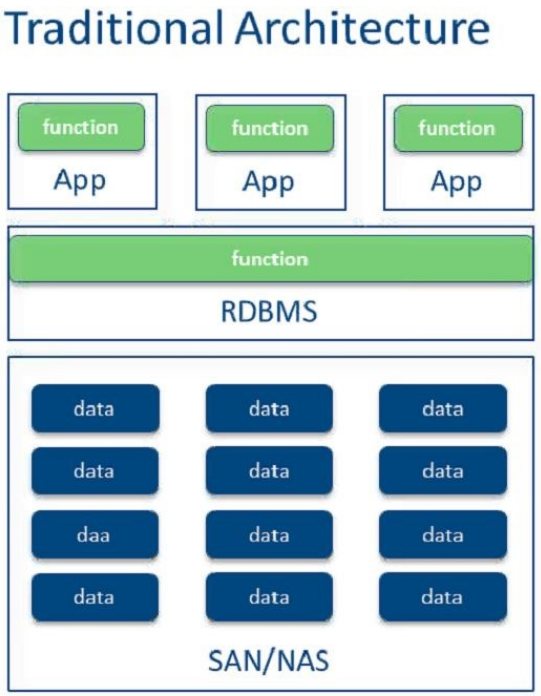
- Week 4 review
- Week 5 contents
  - Chapter 1: Apache Hadoop
  - Chapter 2: Hive and HiveQL, Beeline
  - Chapter 3: Pig Latin, Grunt
  - Chapter 4: SAS-Hadoop, in memory data using Hadoop
- Practical exercises: HDFS, Hive, Pig

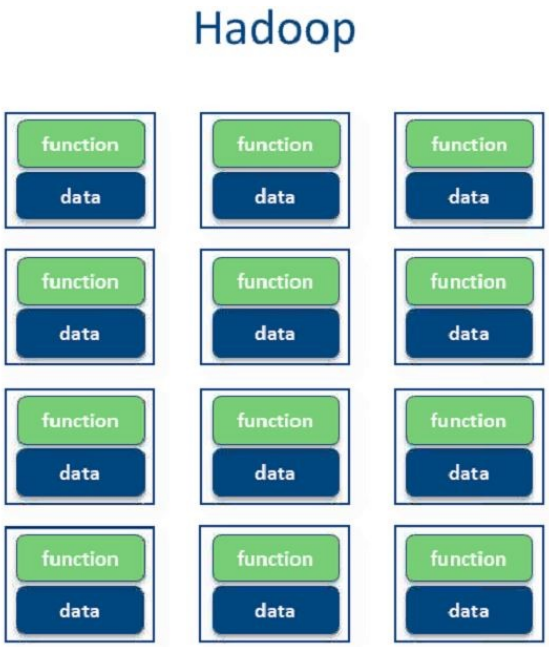
JAMES COOK UNIVERSITY AUSTRALIA

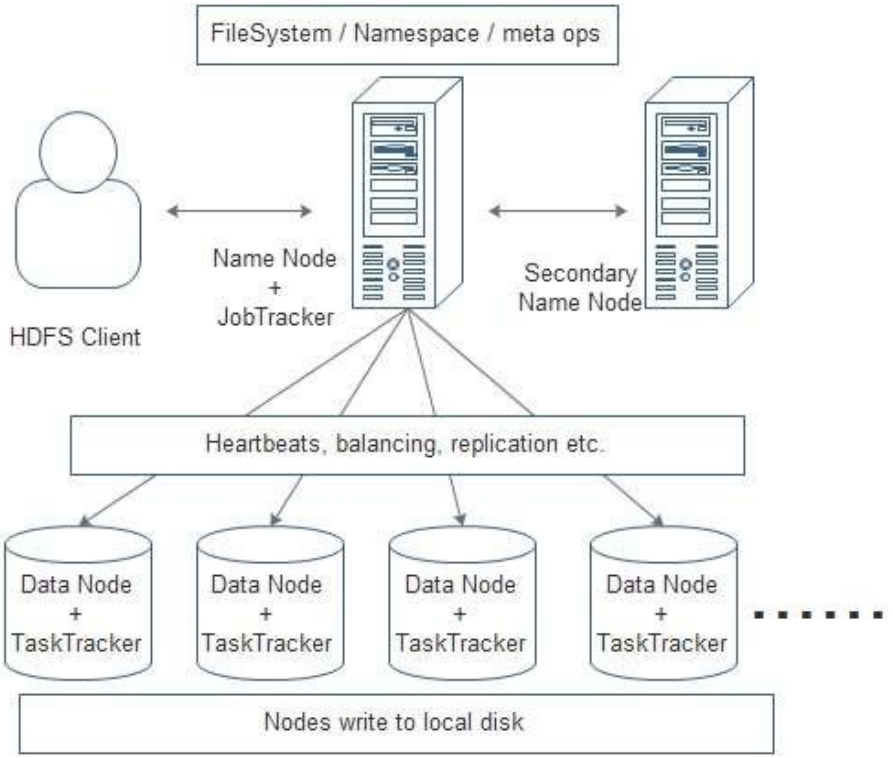
Week 4

- Chapter 1: Introduction to Hadoop in SAS
- Chapter 2: HDFS and MapReduce job through SAS
- Chapter 3: Database (hive) on Hadoop and querying through SAS
- Chapter 4: Accessing data from SAS using LIBNAME
- Chapter 5: Partitioning and Clustering hive tables → HDFS
- Chapter 6: In memory analysis and Hadoop (alternative to SPARK)

Hadoop architecture



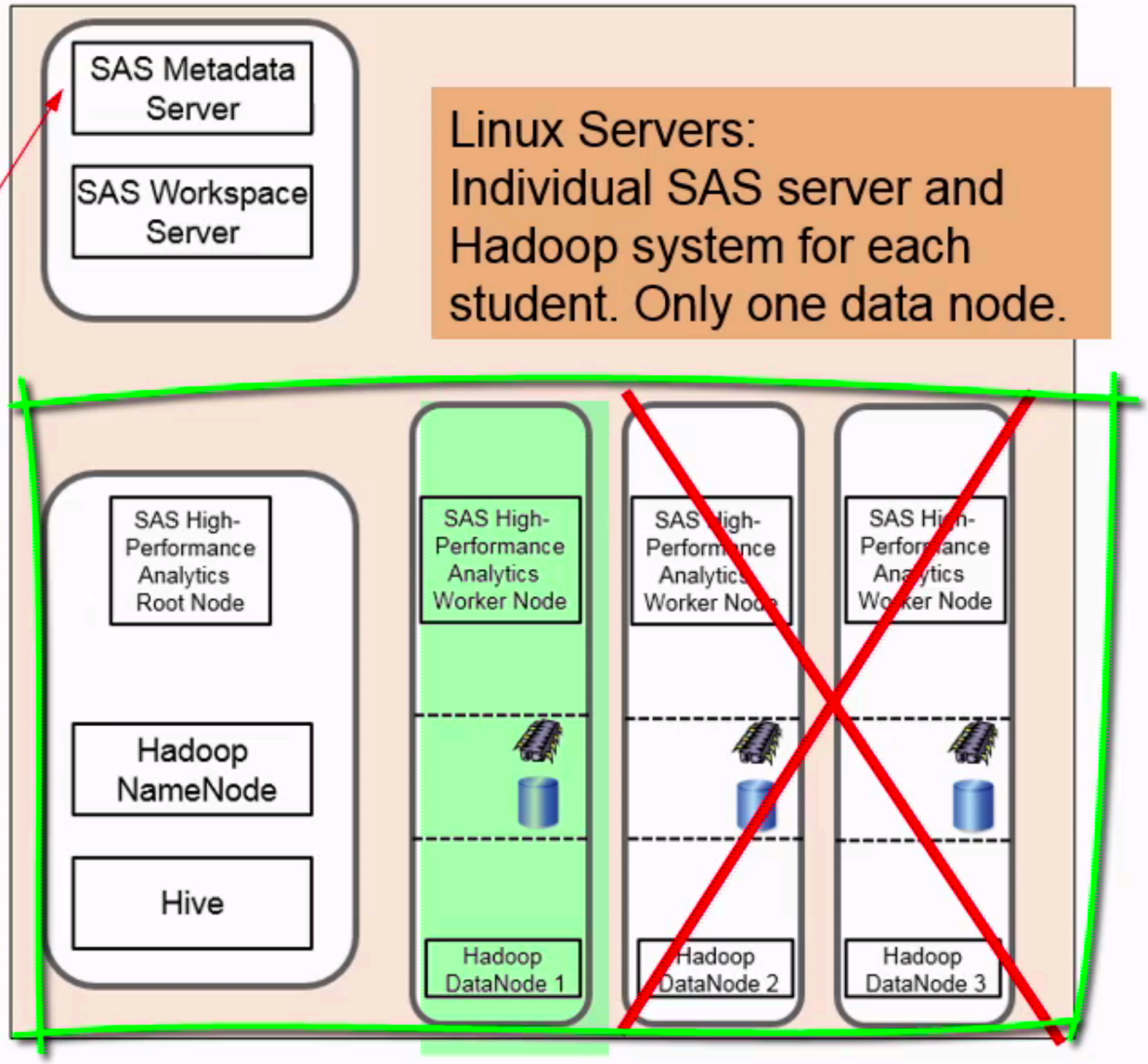




JAMES COOK UNIVERSITY AUSTRALIA

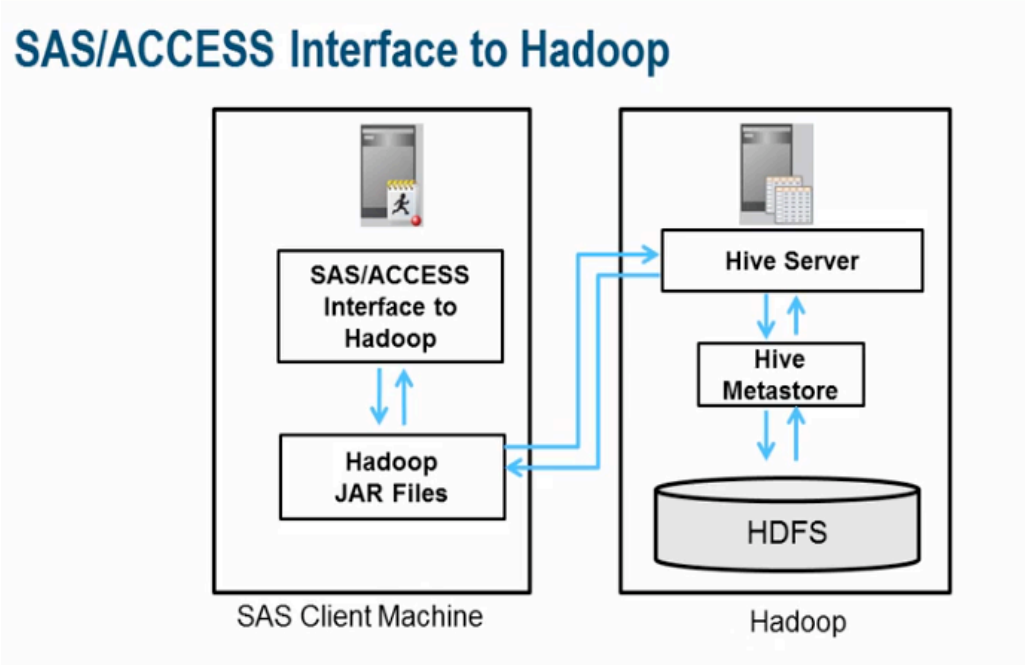
Course Environment in Class

Microsoft Windows Client: Individual machine for each student.



SAS + Hadoop

- HADOOP FILENAME Statement and HADOOP Procedure
- PROC SQL (Executing pass-through statements)
- SAS/ACCESS LIBNAME Statement to interface with Hive
- Hive Tables



Two Types of SAS In-Memory Analytics Products

SAS High-Performance products for

- Statistics
- Data Mining
- Text Mining
- Econometrics
- Forecasting
- Optimization

- procedure language interface similar to corresponding foundation products
- procedures read data in SASHDAT format in parallel from HDFS

SAS LASR Analytic Server products:

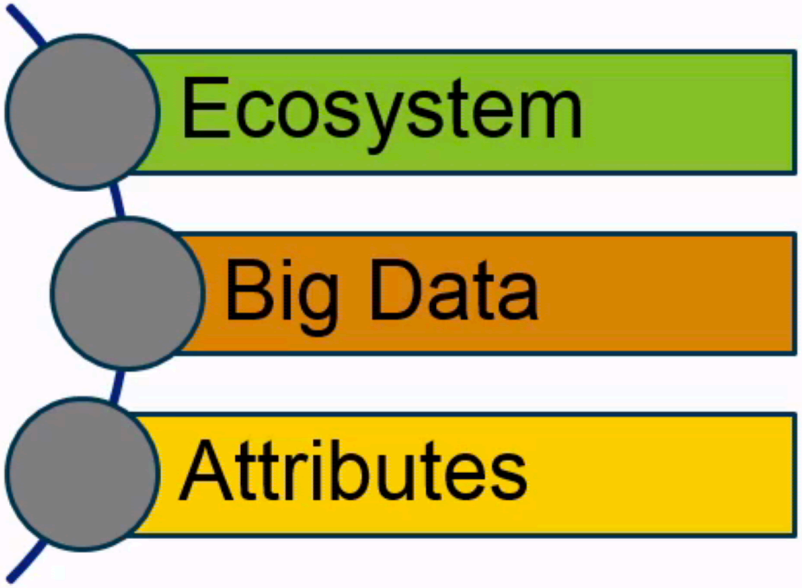
- SAS Visual Analytics
- SAS Visual Statistics
- In-Memory Statistics

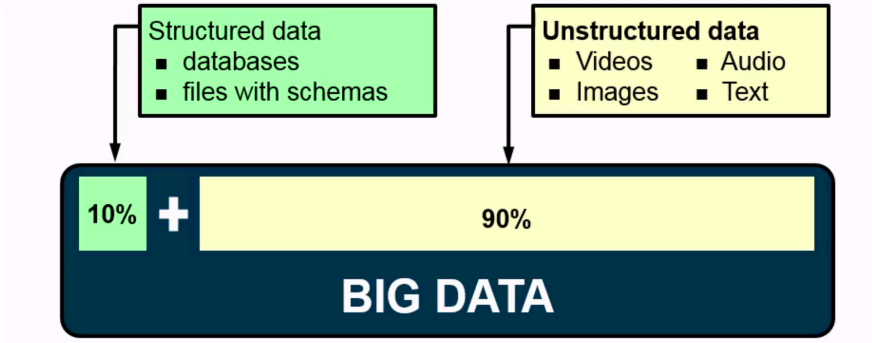
- Visual Analytics and Visual Statistics users create data visualizations with a point and click web application.
- In-Memory Statistics users create analytical results through a language interface.
- Source data is pre-loaded into memory in the LASR Analytic Server grid.

JAMES COOK UNIVERSITY AUSTRALIA

Week 5: Cloudera based HDFS, Hive and Pig

Chapter 1: The Apache Hadoop Project





Why Hadoop

JAMES COOK UNIVERSITY AUSTRALIA

Cheaper: scales to petabytes of more

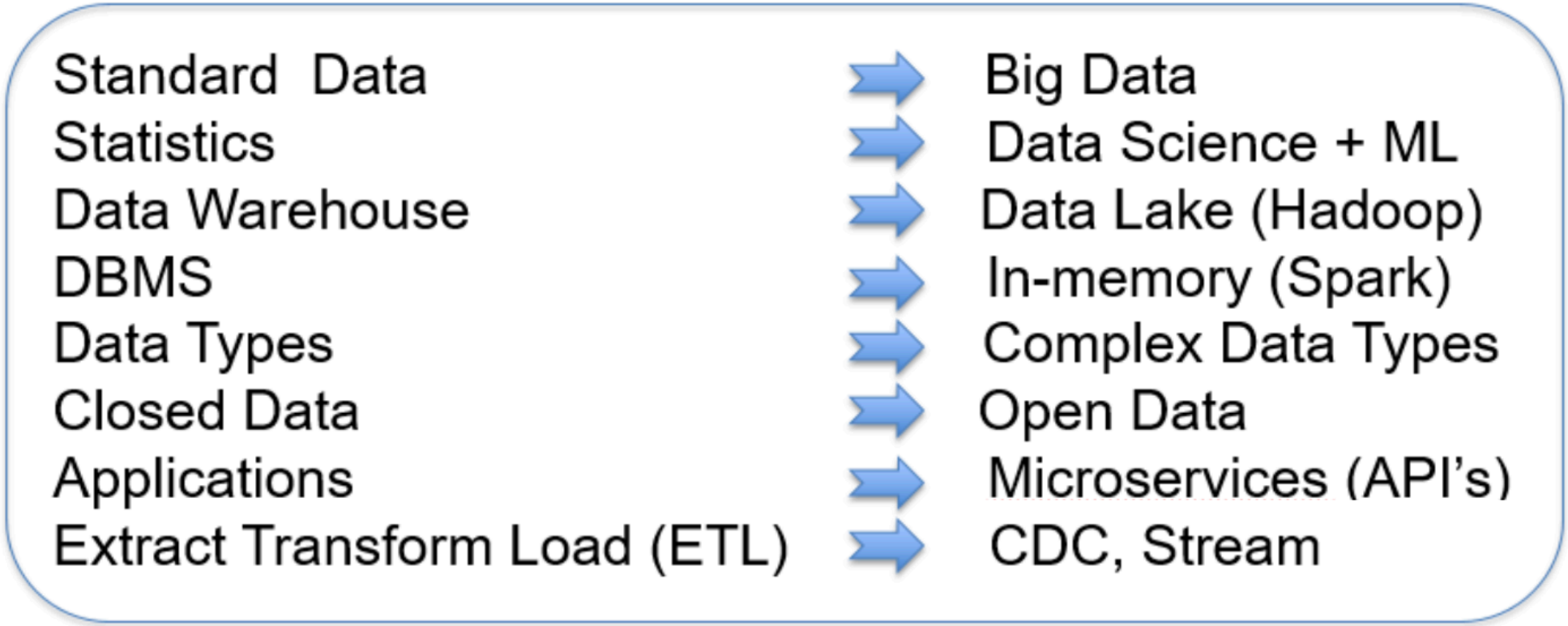
Faster: parallel data processing

Better: suited for particular types of 'Big Data'

Example business problems:

- Risk modelling (insurance company)
- Customer churn analysis (behaviour analysis of customer)
- Recommendation engine (Netflix, eBay, Amazon)
- Ad Targeting (Gmail, social media)
- Transactional Analysis (fraudulent transaction)
- Threat Analysis (counter-terrorism, cyber attack)
- Search Quality (key stroke analysis)

Hadoop- new age of data processing and analysis



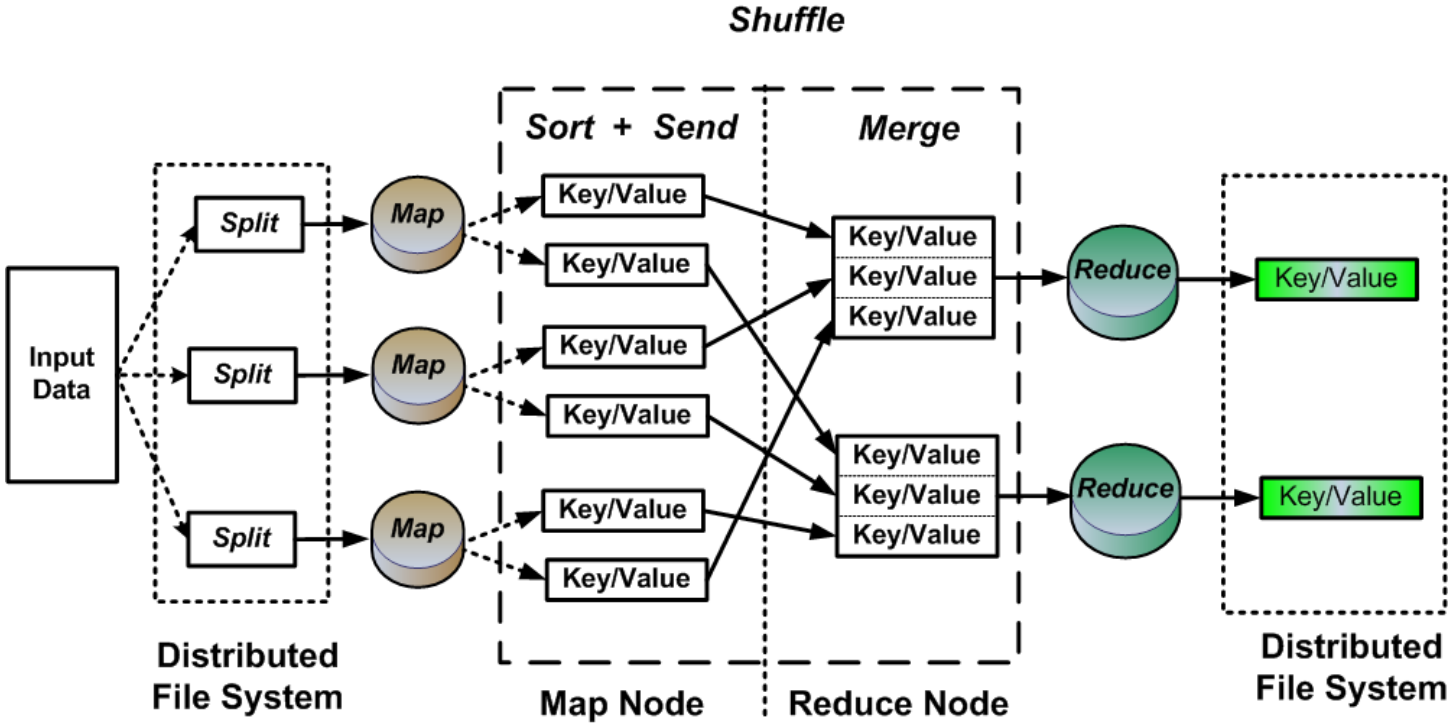
Mapping the technology changes for Analytics, Data Science and Data processing

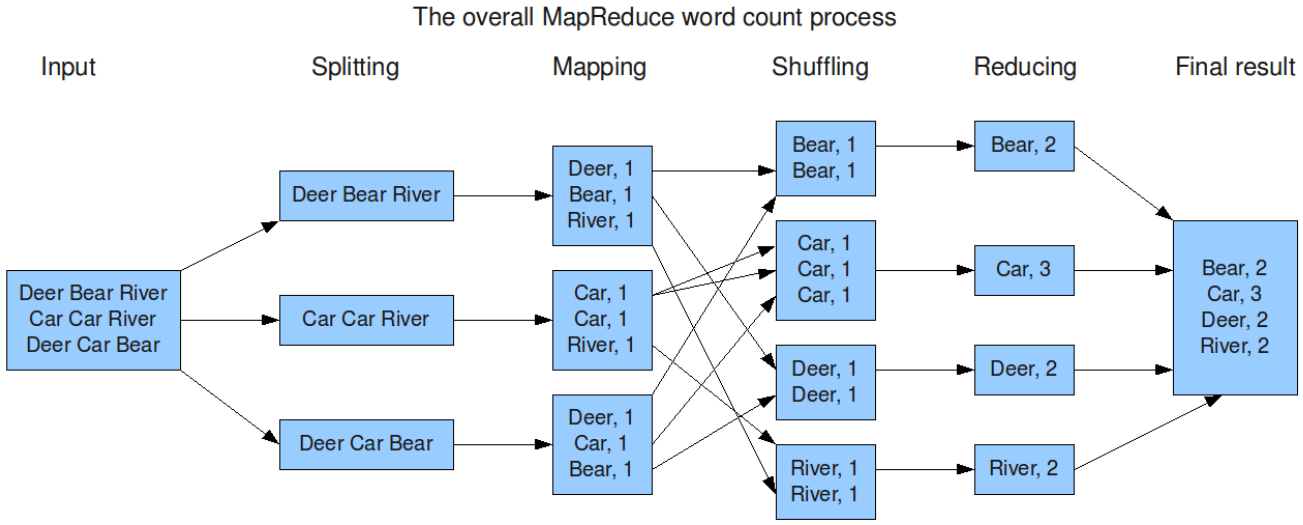
JAMES COOK UNIVERSITY AUSTRALIA

Big Data versus Traditional Technologies

Traditional Data Systems	Big Data Technologies
➢ Rigid data models	➢ Schema free
➢ Weak fault-tolerance architecture	➢ Strong fault-tolerance architecture
➢ Scalability constraints	➢ Highly scalable
➢ Expensive to scale	➢ Economical (1TB ~ 5K)
➢ Limitation for handling unstructured data	➢ Can handle unstructured data
➢ Proprietary hardware and software	➢ Commodity hardware and open-source and proprietary software

Hadoop's computational model

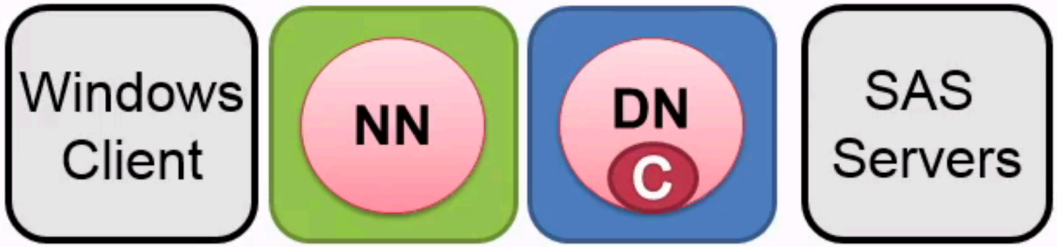




JAMES COOK UNIVERSITY AUSTRALIA

Overview of Course Collection

This class uses a four-machine collection.



- One Windows client. No relation to Hadoop. Contains SAS client software and MySQL database.
- One SAS Servers node running SAS application servers.
- Two-node cluster – Name node (server2.demo.sas.com) with cluster management software and Hue and data node (server3.demo.sas.com).
- No dedicated Client node, as the data node is used as a Client node.

JAMES COOK UNIVERSITY AUSTRALIA

Summary of Big Data Ecosystem

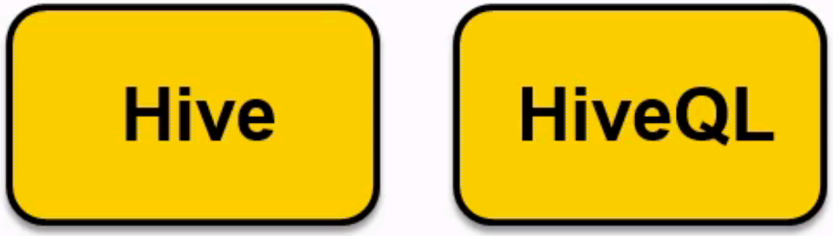
Some Hadoop ecosystem components include the following:

- HDFS (distributed storage) and MapReduce (distributed processing)
  - Job Tracker
  - Task Tracker
- HDFS – Hadoop File System
  - Name node
  - Data node
- Hue – web interface utility to Hadoop
- Sqoop – data movement ***in*** and ***out*** of Hadoop

JAMES COOK UNIVERSITY AUSTRALIA

Chapter 2: Hive and HiveQL

- Overview
- What is Hive?
- HiveQL Data Definition Language
- HiveQL as Data Manipulation Language

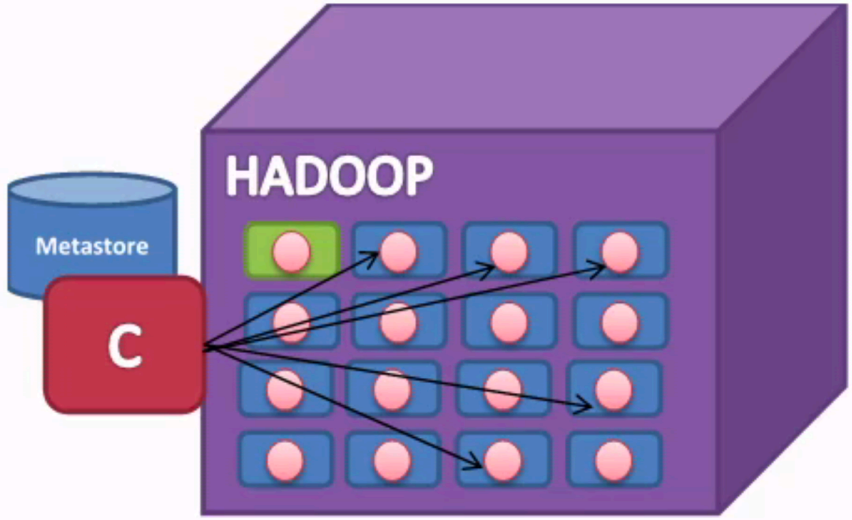


JAMES COOK UNIVERSITY AUSTRALIA

Hive Architecture

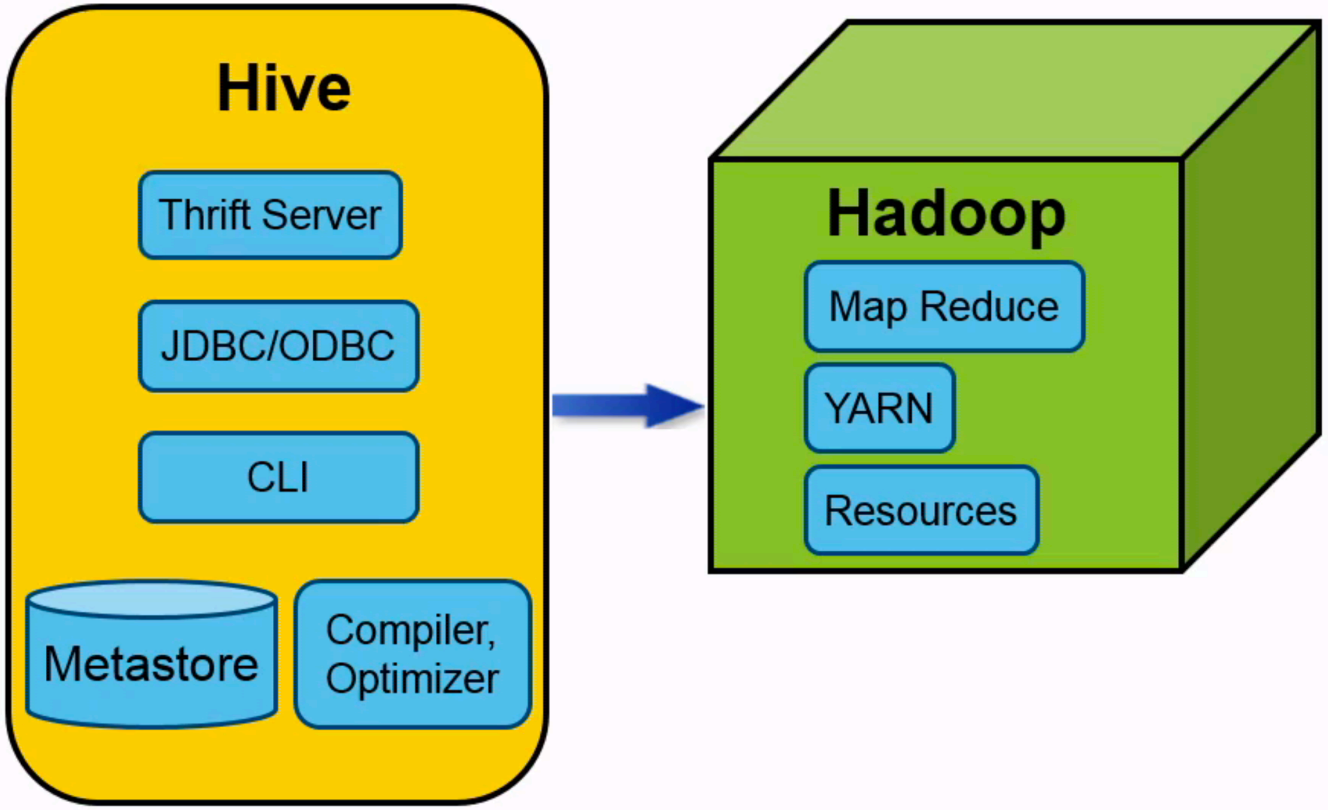
Features of Hive include the following:

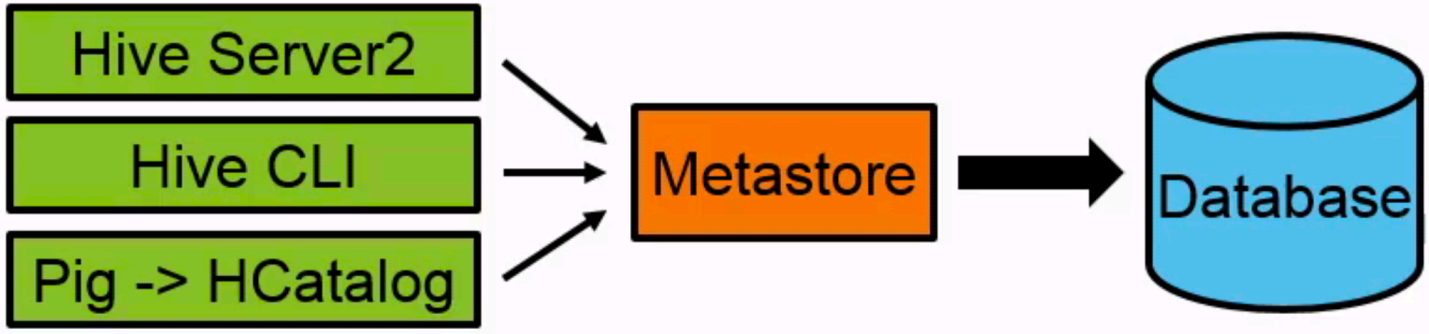
- Hive is a compiler installed on a Client node.
- Hive maintains a metastore of application data outside the Hadoop cluster.
- Hive compiles HiveQL statements into a MapReduce job.
- The MapReduce program is submitted to the Hadoop cluster.



JAMES COOK UNIVERSITY AUSTRALIA

Hive Modules





JAMES COOK UNIVERSITY AUSTRALIA

Summary of Hive DDL

Hive supports the following:

- various primitive data types in Hive (numeric, string, date and time, and miscellaneous)
- complex data types (Array, Map, Struct, UnionType)
- databases and schemas
- managed versus external tables
- creating, dropping, and altering Hive tables
- partitioned tables

JAMES COOK UNIVERSITY AUSTRALIA

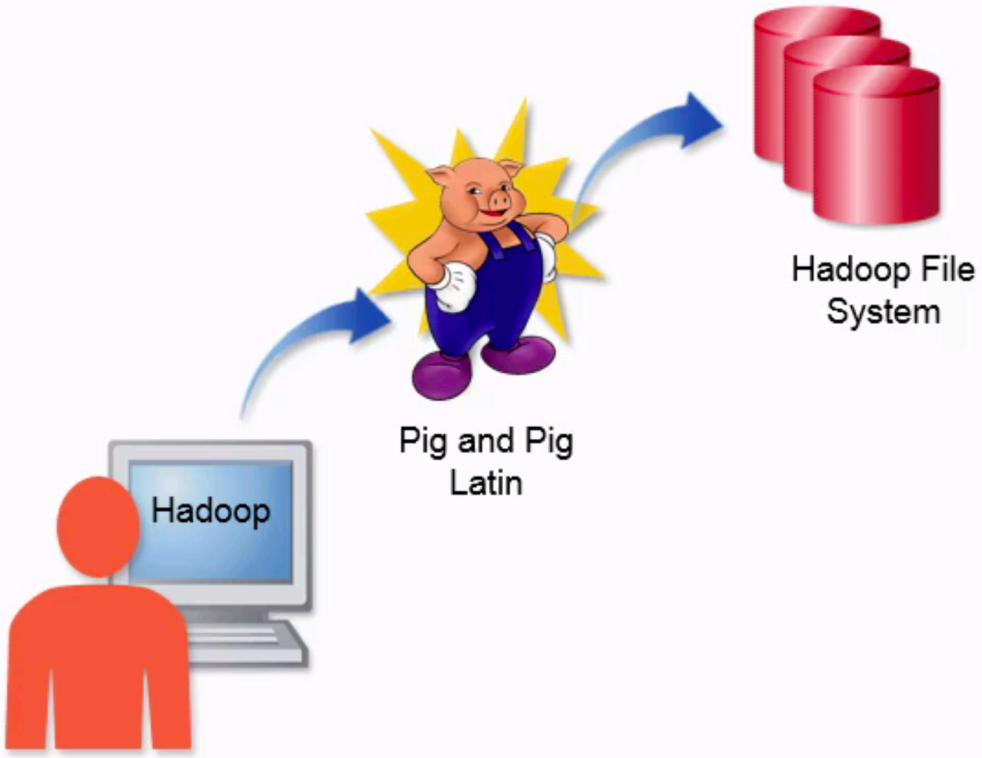
Summary of Hive and HiveQL

A summary of Hive and HiveQL features includes the following:

- Hive manipulation language
- most ANSI-SQL functions available
- SELECT, WHERE, LIMIT, GROUP BY
- Joins (inner, outer)
- loading partitioned tables

JAMES COOK UNIVERSITY AUSTRALIA

Pig and Pig Latin



- Apache Pig
- Pig Programming
- Advanced Operators and Functions
- Recommendations

JAMES COOK UNIVERSITY AUSTRALIA

What Is Pig?

Pig has the following characteristics:

- Pig is the data flow application for writing Hadoop operations with a language called Pig Latin.
- Pig adds a layer of abstraction on top of MapReduce as a high-level language interface to MapReduce instead of using Java.
- Pig provides a SQL-like interface to process data on Hadoop. This helps the programmer focus on business logic and helps increase productivity.
- Pig was initially developed at Yahoo. Pig enables people using Apache Hadoop to focus more on analyzing large data sets and to spend less time writing MapReduce programs in Java.

JAMES COOK UNIVERSITY AUSTRALIA

Pig Components

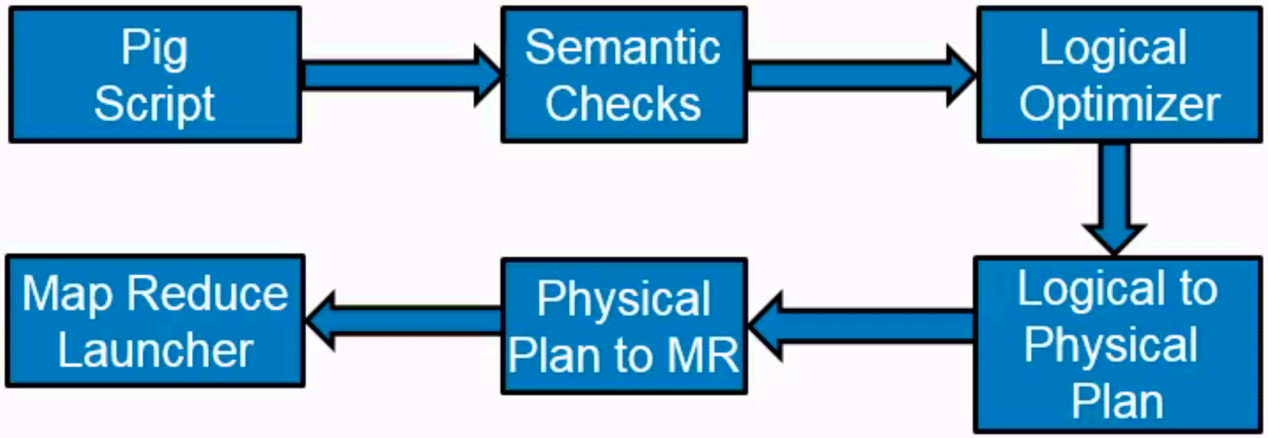
Pig consists of the following components:

- **Pig Latin:** Command-based language, designed specifically for data transformation and flow expression.
- **Grunt:** Command-line tool for running and debugging Pig statements. This is also known as *interactive mode* because a user enters Pig commands and statements interactively on the command line.

How Pig Works

Pig operations include the following:

- The Pig Compiler performs a syntax check.
- A logical optimizer combines various Pig language steps into unified tasks.
- A logical plan is converted into a physical plan.
- A physical plan is further converted into MapReduce.
- The MapReduce job is launched on the cluster.



JAMES COOK UNIVERSITY AUSTRALIA

Comparison of SQL versus Pig

```
create table t2 as
   select col1, col2
      from table1
   where col1 > 10
      and col2 = col1
   limit 10;
```

```
T = LOAD 'data' Using PigStorage('\t')
   AS (col1:int, col2:int, col3:int);
T0 = FOREACH T GENERATE col1, col2;
T1 = FILTER T0
          BY col1 > 10
         AND col2 == col1;
T2 = LIMIT T1 10;
STORE T2 INTO 'HDFS-DIR';
```

- T, T0, T1, T2 are referred as data “aliases.”

Local Mode versus Map Reduce Mode

Two modes of Pig job execution are shown below.

Mode	What is it?	Purpose?	Where?	How?
Local	Local machine with local resources	Good for experimentation, development, and prototyping	Single JVM	$pig –x local
MapReduce	Default mode running on a Hadoop cluster	Production jobs	Hadoop cluster	$pig –x mapreduce

JAMES COOK UNIVERSITY AUSTRALIA

Apache Pig Data Types

Supported data types in Apache Pig include the following:

Int	Signed 32-bit integer
Long	Signed 64-bit integer
Float	32-bit floating point
Double	64-bit floating point
Chararray	Character array (string) in Unicode UTF-8 format
Boolean	Boolean value
Bytearray	Byte array (blob)
Datetime	For dates and time
Tuple	An ordered set of fields
Bag	A collection of tuples
Null	Implemented using the SQL definition of null as unknown or non-existent. Nulls can occur naturally in data or can be the result of an operation

JAMES COOK UNIVERSITY AUSTRALIA

Field, Tuple, Databag, Relations

Pig Latin types include the following:

- Pig Latin statements work on relations.
- Relations are unordered.
- Relations are referred to by *alias*.
- Aliases are assigned as part of a Pig Latin statement.

Type	Description	Example
Field	Piece of data	10
Tuple	An ordered set of fields	(1,2,3)
Bag	Collection of tuples	{(1,2,3),(4,5,6)}
Relation	An ***outer*** bag	Example: {(1,2,3),(4,5,6)} A bag containing two tuples Example: {{(1,2,3)},{(4,5,6)}} An outer bag with two inner bags that contain one tuple each

JAMES COOK UNIVERSITY AUSTRALIA

Expressions

Examples of Pig expressions:

- **Arithmetic** expression

```
X = Group A BY f2 + f3;
```

- **String** expression

```
X = FOREACH A GENERATE CONCAT(f2,f3);
```

- **Boolean** expression

```
X = FILTER A by (F1==8) OR (NOT (f2 + f3 > f1));
```

- **STAR** expression

```
A = LOAD 'data' USING PIGSTORAGE() AS
     (name:chararray, age:int);
B = FOREACH A GENERATE COUNT(*);
```

JAMES COOK UNIVERSITY AUSTRALIA

Flatten Operator

Flattening a bag results in one row for each tuple in the bag:

```
B = FOREACH A GENERATE flatten($0);
```

A:

({(b,c),(d,e)})

B:

(b,c)
(d,e)

When other fields are projected, they are repeated in each row.

```
B = FOREACH A GENERATE $0, flatten ($1);
```

A:

(a, {(b,c),(d,e)})

B:

(a,b,c)
(a,d,e)

JAMES COOK UNIVERSITY AUSTRALIA

Keyword: PigStorage ... AS

The PigStorage function parses the lines of delimited input files into separate fields. The default delimiter is a tab.

Usage	Arguments	Behavior
`PigStorage()`	None	Default "\t" is assumed
`PigStorage('\t')`	Tab character	Tab-delimited
`PigStorage(',')`	Comma character	Comma-delimited
`PigStorage(',', '-schema')`	Delimiter, Schema	Reads or stores the schema of the relationship using a hidden JSON file
`PigStorage(',', '-noschema')`	Delimiter, NoSchema	Ignores any associated schema
`PigStorage(',', '-tagsource')`		Appends the input source file path to the end of each tuple. Ensure "pig.splitCombination=false"

Language Manual References

IMPALA

https://www.cloudera.com/documentation/enterprise/5-5-x/topics/impala_string_functions.html
https://www.cloudera.com/documentation/enterprise/5-6-x/topics/impala_analytic_functions.html

HIVE

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Select

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+GroupBy

Cloudera

http://tiny.cloudera.com/dac10ab
http://tiny.cloudera.com/dac10a

https://www.gartner.com/doc/reprints?id=1-1M5EDSWC&ct=190712&st=sb