

Capítulo 5: Arquitectura de Hadoop

Capítulo 6: Componentes de Hadoop

Capítulo 7: Administración de Hadoop



DAT
División de Alta Tecnología

6

Componentes de Hadoop

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.




Objetivos

Al finalizar el capítulo, el alumno logrará:

- Crear Jobs MapReduce.
- Utilizar Pig
- Utilizar Hive
- Utilizar Flume
- Utilizar Sqoop
- Utilizar Oozie

6 - 2

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Agenda

- MapReduce
- Pig y Hive
- Flume y Sqoop
- Oozie

6 - 3

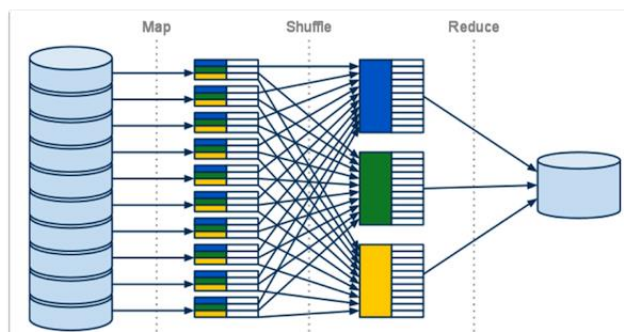
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Map Reduce

Estrategia divide y vencerás:

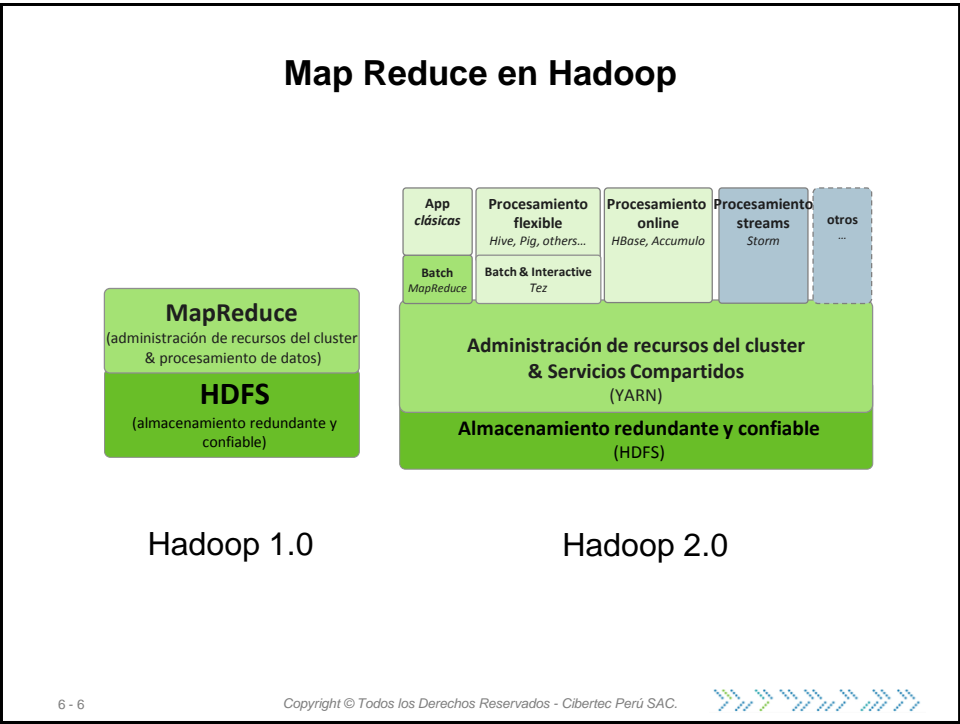
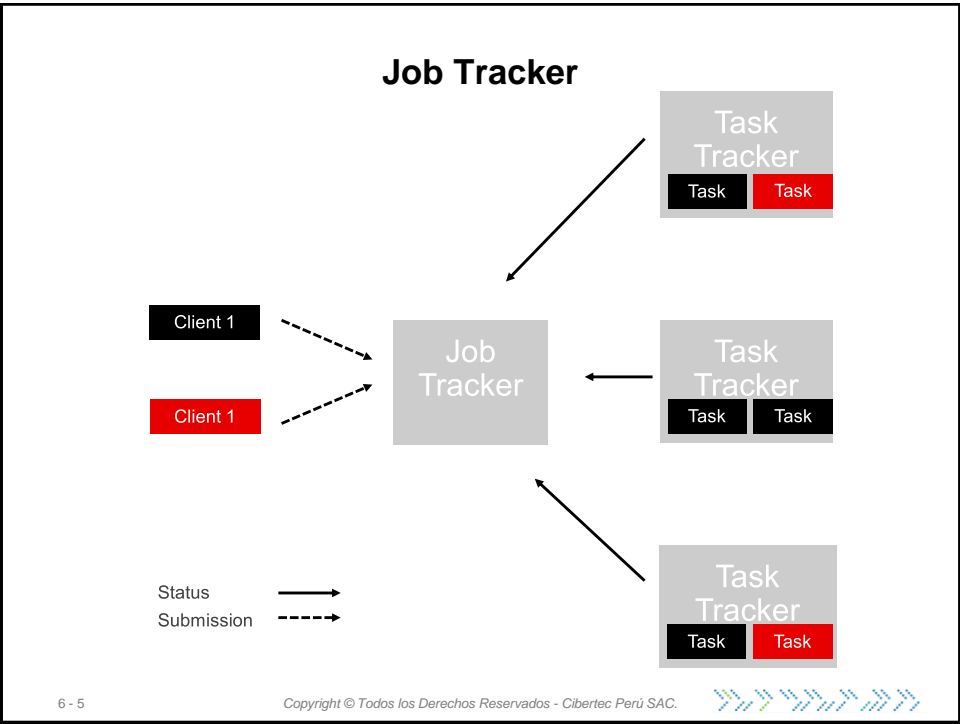
- Map() – dividir en problemas más pequeños
- Reduce() – combinar los resultados



6 - 4

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.

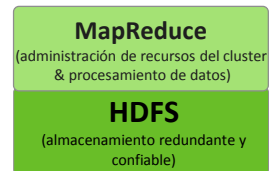




Hadoop 1.0 con Map Reduce

Hadoop 1.0

- Componentes
 - HDFS
 - Map Reduce
- Aplicaciones en Batch
- Permite ejecutar únicamente Map Reduce



6 - 7

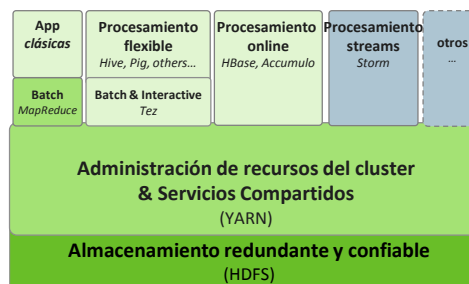
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Hadoop 2.0 con Map Reduce

Hadoop 2.0

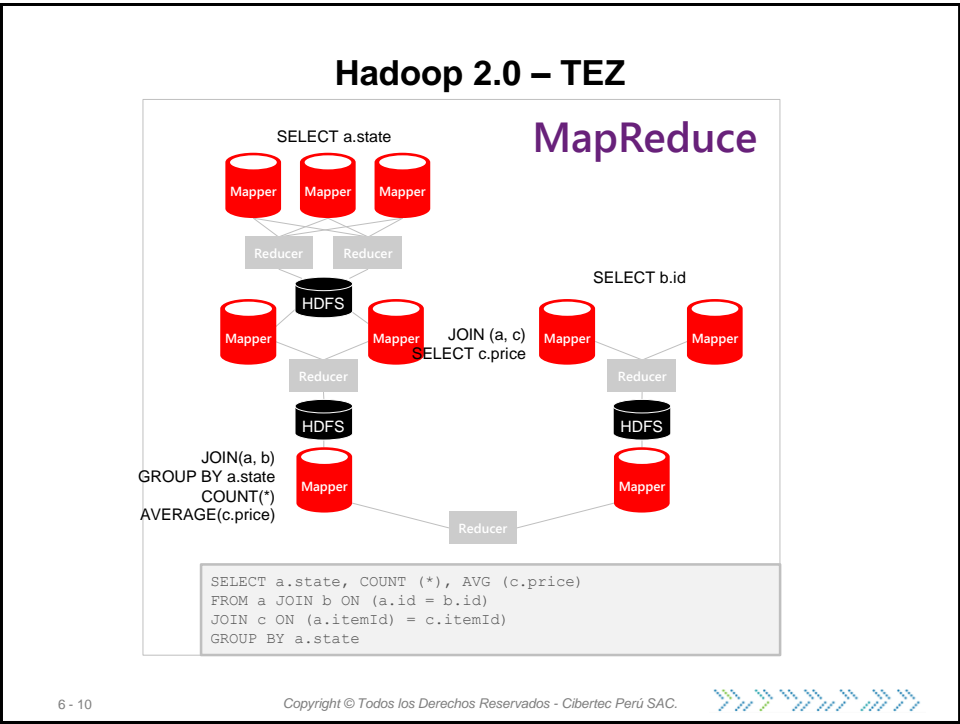
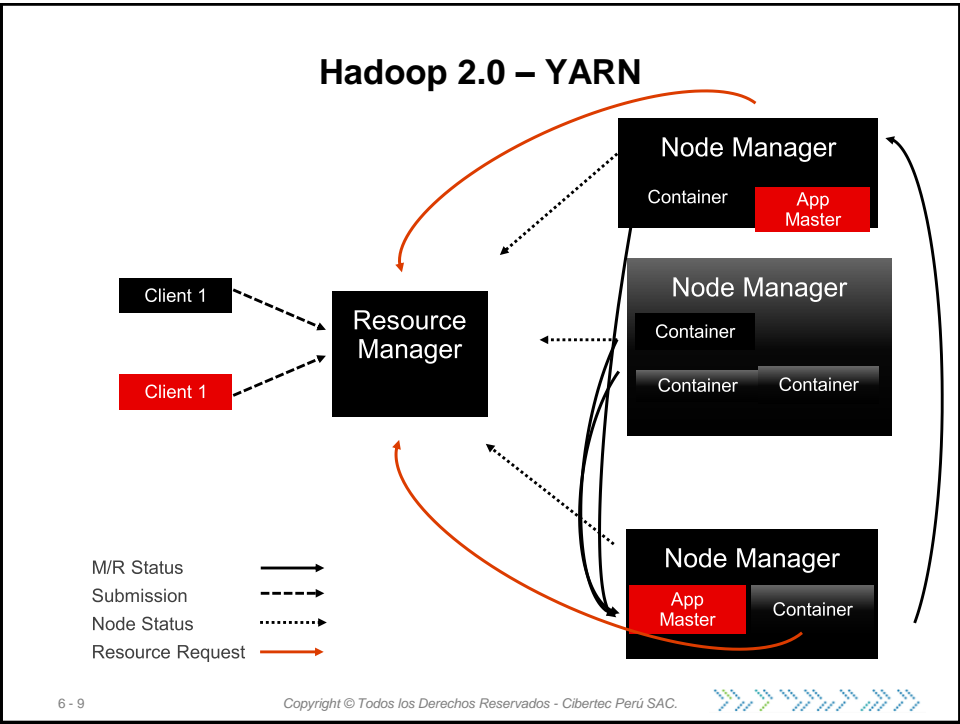
- Componentes
 - YARN
 - Tez
- Aplicaciones en batch, interactivas
- Permite ejecutar estrategias diferentes a Map Reduce



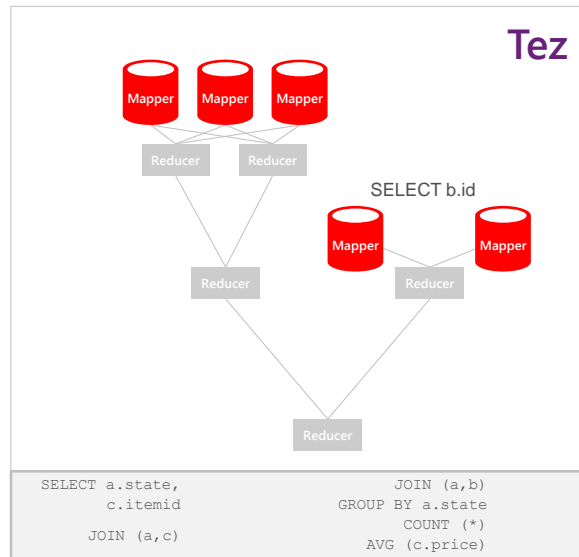
6 - 8

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.





Hadoop 2.0 – TEZ



6 - 11

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Pig

- Permite analizar grandes conjuntos de datos semiestructurados y no estructurados
- Genera jobs de Map Reduce
- Principalmente enfocada en ETL
- Se generan transformaciones mediante “relaciones” de datos
- No requiere esquema. Las relaciones se cargan utilizando esquema en lectura
- Utiliza Pig Latin como lenguaje
- Ejecutar de forma interactiva (consola / grunt) o en modo batch (script)



6 - 12

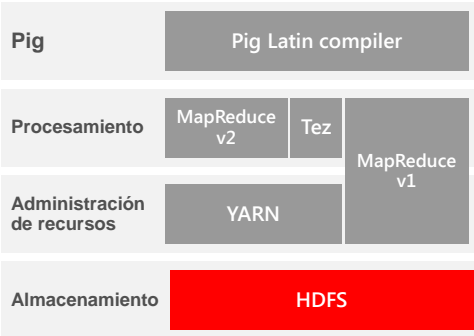
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Pig

Fases de ejecución

- Parsing
- Optimización
- Plan de ejecución
 - Lógico
 - Físico (Map Reduce) STORE / DUMP



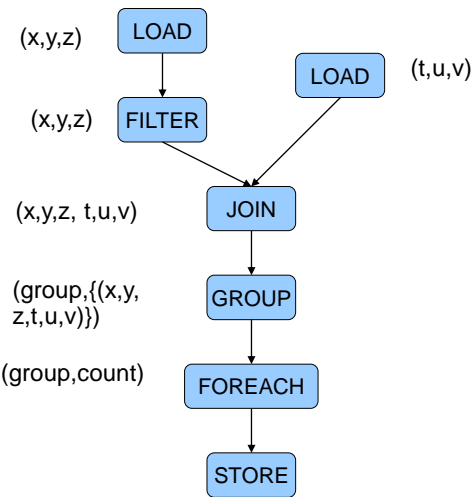
6 - 13

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Pig

```
A=LOAD 'file1' AS (x, y, z);
B=LOAD 'file2' AS (t, u, v);
C=FILTER A by y > 0;
D=JOIN C BY x, B BY u;
E=GROUP D BY z;
F=FOREACH E GENERATE
  group, COUNT(D);
STORE F INTO 'output';
```



6 - 14

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Pig - Conceptos

Conceptos:

- **Atom:** Valor simple almacenado como string
- **Tuple:** Registro de datos que consiste de una secuencia de campos (eq: fila) “()”
- **Bag/Relation:** Conjunto de tuplas “{}”
- Las tuplas pueden tener tipos de datos diferentes
- **Map:** mapeo de llaves utilizando una estrategia de hash “->”
 - La llave es un string (atom) y el valor puede ser de cualquier tipo de dato

```
{
  Name-> 'Bicycle',
  Price -> 105,
  Parts -> {
    (1, 'Wheel', 2, 10.00 ),
    (2, 'Chain', 1, 15.00 ),
    (3, 'Handlebars', 1, 3.00),
    ...
  };
  Name -> 'Tricycle',
  Price -> $55,
  Parts -> {
    (1, 'Wheel', 3),
    (3, 'Handlebars', 1),
    ...
  }
}
```

6 - 15

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Pig - Expresiones

$t = ('alice', \{ ('lakers', 1), ('iPod', 2) \}, ['age' \rightarrow 20])$ <p>Let fields of tuple t be called f1, f2, f3</p>		
Expression Type	Example	Value for t
Constant	'bob'	Independent of t
Field by position	\$0	'alice'
Field by name	f3	'age' → 20
Projection	f2.\$0	{ ('lakers'), ('iPod') }
Map Lookup	f3#'age'	20
Function Evaluation	SUM(f2.\$1)	1 + 2 = 3
Conditional Expression	f3#'age'>18? 'adult': 'minor'	'adult'
Flattening	FLATTEN(f2)	'lakers', 1 'iPod', 2

6 - 16

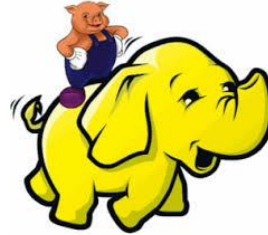
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Pig Latin

Instrucciones

- LOAD
- FILTER
- FOREACH GENERATE
- ORDER
- JOIN
- GROUP
- LIMIT
- FLATTEN
- STORE
- DUMP



Agregaciones

- Count, Avg, Sum, Max, Min

6 - 17

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Pig Latin

FOREACH (proyección)

Aplica procesamiento a cada tupla de la relación

DUMP alias1;

(1,2,3) (4,2,1) (8,3,4) (4,3,3) (7,2,5) (8,4,3)

alias2 = FOREACH alias1 GENERATE col1, col2;

DUMP alias2;

(1,2) (4,2) (8,3) (4,3) (7,2) (8,4)

6 - 18

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Pig Latin

FILTER (reducir tuplas)

```
DUMP alias1;  
(1,2,3) (4,2,1) (8,3,4) (4,3,3) (7,2,5) (8,4,3)  
alias2 = FILTER alias1 BY (col1 == 8) OR (NOT (col2+col3 > col1));
```

```
DUMP alias2;  
(4,2,1) (8,3,4) (7,2,5) (8,4,3)
```

6 - 19

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Pig Latin

GROUP

```
DUMP alias1;  
(John,18,4.0F) (Mary,19,3.8F) (Bill,20,3.9F) (Joe,18,3.8F)  
alias2 = GROUP alias1 BY col2;
```

```
DUMP alias2;  
(18,{{John,18,4.0F},{Joe,18,3.8F}})  
(19,{{Mary,19,3.8F}})  
(20,{{Bill,20,3.9F}})
```

6 - 20

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Pig Latin

JOIN (Inner / Outer)

DUMP Alias1;

(1,2,3) (4,2,1) (8,3,4) (4,3,3) (7,2,5) (8,4,3)

DUMP Alias2;

(2,4) (8,9)(1,3)(2,7)(2,9)(4,6)(4,9)

Alias3 = JOIN Alias1 BY Col1, Alias2 BY Col1;

Dump Alias3;

(1,2,3,1,3) (4,2,1,4,6)(4,3,3,4,6)(4,2,1,4,9)(4,3,3,4,9)(8,3,4,8,9)

(8,4,3,8,9)

6 - 21

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Pig Latin

Se puede hacer referencia a los campos mediante la posición (\$0, \$1, \$2) o el alias

```
students = LOAD 'student.txt' USING PigStorage() AS  
(name:chararray, age:int, gpa:float);
```

```
studentname = Foreach students Generate $1 as studentname;
```

6 - 22

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Pig Latin

- LOAD para leer datos en relaciones (bag)

```
Loaders (PigStorage, TextLoader, ... )
var = LOAD 'employees.txt';
var = LOAD 'employees.txt' AS (id, name, salary);
var = LOAD 'employees.txt' using PigStorage()
AS (id, name, salary);
```

- Transformaciones
- STORE (Archivo en HDFS)


```
grunt> STORE processed INTO 'processed_txt';
```
- DUMP (Pantalla)


```
grunt> DUMP processed;
```

6 - 23

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Hive

- Lenguaje basado en SQL llamado HiveQL para consultar los datos
- Permite consultar, agregar y analizar los datos almacenados en HDFS
- Se utiliza para consumir datos de Hadoop en herramientas de BI
- Genera ejecuciones de Map Reduce (MR 1.0, Tez)
- Se puede ejecutar de forma interactiva (consola) o en modo batch (script)



6 - 24

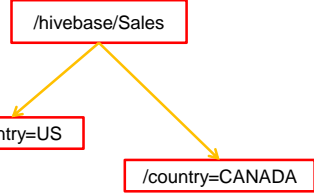
Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Hive

Tablas

- Asociada a un directorio en HDFS
- Datos se almacenan como archivos en el directorio
- Columnas con tipo de dato (int, float, string, boolean)



Particiones

- Determinar la distribución de los datos en subdirectorios

Buckets

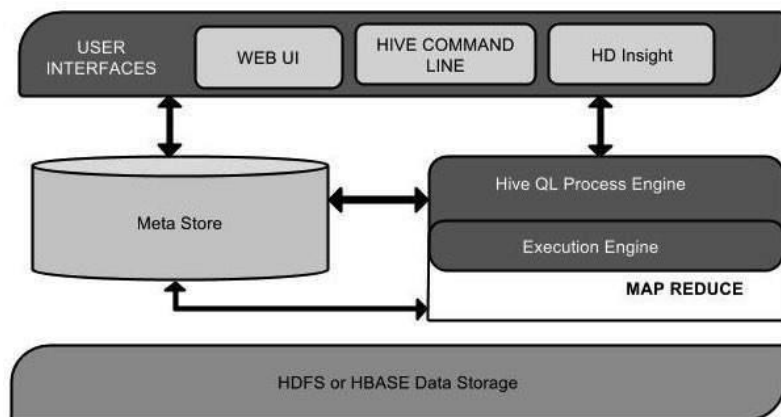
- Divide una partición en buckets mediante en una función de hash
- Cada bucket se almacena como un archivo en el directorio de la partición

6 - 25

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Arquitectura Hive



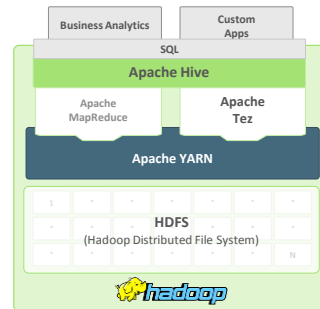
6 - 26

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Hive QL

- Servicio de meta data que proyecta esquemas tabulares sobre carpetas en HDFS
- DDL
 - CREATE TABLE, ALTER TABLE, SHOW TABLE, DESCRIBE
- DML
 - LOAD TABLE, INSERT, UPDATE, DELETE
- QUERY
 - SELECT, GROUP BY, JOIN



6 - 27

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Hive- Create Table

Tabla Interna

```
CREATE TABLE table1
(col1 STRING,
 col2 INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

```
CREATE TABLE table2
(col1 STRING,
 col2 INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/data/table2';
```

Tabla Externa

```
CREATE EXTERNAL TABLE table3
(col1 STRING,
 col2 INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/data/table3';
```

6 - 28

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Hive- Load

LOAD (Copia los datos a la carpeta indicada)

```
LOAD DATA LOCAL INPATH '/data/source' INTO TABLE  
MyTable;
```

INSERT (inserta de una tabla a otra)

```
FROM StagingTable INSERT INTO TABLE MyTable  
SELECT Col1, Col2;
```

```
INSERT OVERWRITE DIRECTORY '/tmp/hdfs_out' SELECT  
* FROM sample WHERE ds='2012-02-24';
```

6 - 29

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Hive- Select

```
SELECT Col1, SUM(Col2) AS TotalCol2  
FROM MyTable  
WHERE Col1 >= '2013-06-01' AND Col1 <= '2013-06-30'  
GROUP BY Col1 ORDER BY Col1;
```

```
SELECT MAX(foo) FROM sample;
```

```
SELECT ds, COUNT(*), SUM(foo)  
FROM sample  
GROUP BY ds;
```

```
SELECT * FROM customer c  
JOIN order_cust o ON (c.id=o.cus_id);  
ORDER BY c.id LIMIT 10
```

6 - 30

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Flume

Apache Flume es un mecanismo de ingestión de herramienta/servicio/datos para recolectar agregados y transportar grandes cantidades de datos de transmisión como archivos de registro, eventos (etc ...) de diversas fuentes a un almacén de datos centralizado.

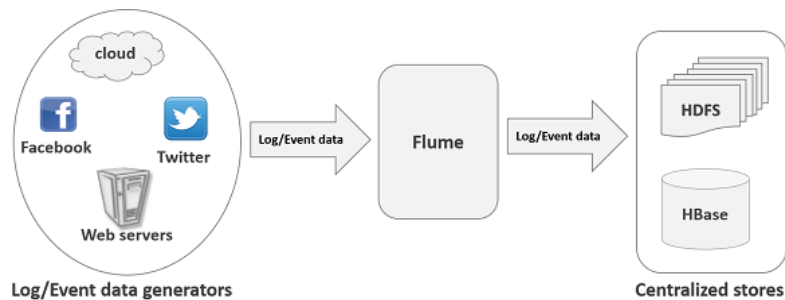
Flume es una herramienta altamente confiable, distribuida y configurable. Está diseñado principalmente para copiar datos de transmisión (datos de registro) desde varios servidores web a HDFS.

6 - 31

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Flume

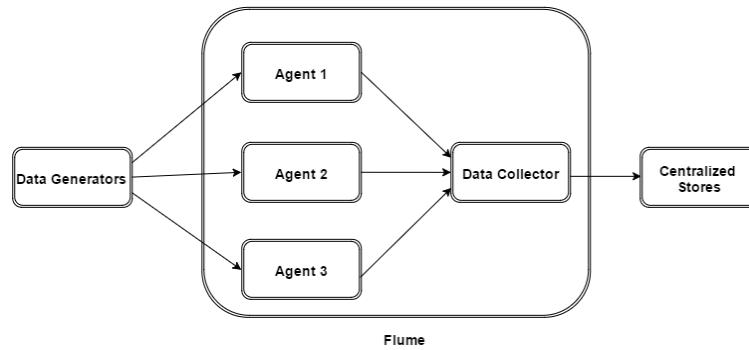


6 - 32

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Arquitectura Flume



6 - 33

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Sqoop

- Integración con bases de datos.
- Utiliza JDBC para realizar la conexión.
- Interfaz de línea de comandos para realizar transferencia de datos.
- Soporta cargas incrementales.
- Import: Cargar datos a Hadoop.
- Export: Extraer datos de Hadoop a base de datos.
- Se integra con Oozie para permitir programación y automatización.



6 - 34

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Sqoop

```
$ sqoop import --connect  
jdbc:mysql://localhost/DB_NAME --table TABLE_NAME --  
username USER_NAME --password PASSWORD
```

```
$ sqoop export --connect jdbc:mysql://SERVER/DB_NAME  
--table TARGET_TABLE_NAME --username USER_NAME --  
password PASSWORD --export-dir EXPORT_DIR
```

6 - 35

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Oozie

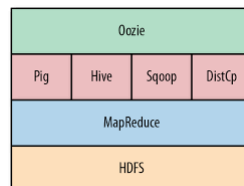
Oozie es un programador para administrar los jobs de Apache Hadoop

Oozie Workflow Document

- Archivo XML con las acciones

Archivos de script

- Archivos utilizados por las acciones
(ej: HiveQL query file)



6 - 36

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Oozie

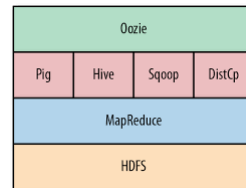
Oozie es un programador para administrar los jobs de Apache Hadoop

Oozie Workflow Document

- Archivo XML con las acciones

Archivos de script

- Archivos utilizados por las acciones (ej: HiveQL query file)



6 - 37

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Apache Oozie

```
<workflow-app xmlns="uri:oozie:workflow:0.2" name="MyWorkflow">
  <start to="FirstAction"/>
  <action name="FirstAction">
    <hive xmlns="uri:oozie:hive-action:0.2">
      <script>CreateTable.q</script>
      <param>TABLE_NAME=${tableName}</param>
      <param>LOCATION=${tableFolder}</param>
    </hive>
    <ok to="SecondAction"/>
    <error to="fail"/>
  </action>
  <action name="SecondAction">
    ...
  </action>
  <kill name="fail">
    <message>Workflow failed, error message[${wf:errorMessage(wf:lastErrorNode())}]</message>
  </kill>
</end name="end"/>
</workflow-app>
```

6 - 38

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Ejercicio N° 6: Componentes de Hadoop

Al finalizar la tarea, el alumno logrará:

- Aprender como trabajar con los diversos componentes de Hadoop, tales como: Sqoop, Hive, Flume, Pig

6 - 39

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Tarea 6: Curso Hive, Sqoop y Flume

Al finalizar la tarea, el alumno logrará:

- Aprenderá que es Hive un sistema de almacenamiento de datos para Hadoop que facilita el resumen de datos, consultas ad-hoc y el análisis de grandes conjuntos de datos almacenados en sistemas de archivos compatibles con Hadoop.
- Aprenderá como importar o cargar datos en HDFS desde fuentes de datos comunes, como bases de datos relacionales, almacenes de datos, registros del servidor web.

6 - 40

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Lecturas adicionales

Se sugiere revisar los siguientes enlaces para profundizar en los conceptos tratados en el presente capítulo:

- a) Procesos MapReduce

6 - 41

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



Resumen

En este capítulo, hemos aprendido como trabajar con los Jobs de MapReduce, a utilizar Pig con el lenguaje Pig Latin, a realizar consultas con Hive, a implementar Flume, Sqoop y Oozie.

6 - 42

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.

