Revision 1

Fundamentals of Big Data

1. Hadoop uses\_\_\_\_ to allow authentication access for services.

2. \_\_\_ is a special token that acts as a pass at name node to request the service of

a datanode

3. \_\_\_\_manages filesystem tree, metadata and directories in the tree in a HDFS

cluster

4. Output of map phase in a simple MapReduce data Processing model could be

5. Output of Reduce phase in a simple MapReduce data Processing model could be

6. \_\_\_\_\_ which hashes a record&#39;s key to determine the partition by a reduce task

7. Input split can be computed by using the formula \_\_\_\_-

8.

&quot;On the top of the Crumpetty Tree

The Quangle Wangle sat,

But his face you could not see,

On account of his Beaver Hat.&quot;

if this is the input of MR task what can be the key for the line 1

9. &quot;On the top of the Crumpetty Tree

The Quangle Wangle sat,

But his face you could not see,

On account of his Beaver Hat.&quot;

if this is the input of MR task what can be the key for the line 2

10. &quot;On the top of the Crumpetty Tree

The Quangle Wangle sat,

But his face you could not see,

On account of his Beaver Hat.&quot;

if this is the input of MR task what can be the key for the line 3

11. &quot;On the top of the Crumpetty Tree

The Quangle Wangle sat,

But his face you could not see,

On account of his Beaver Hat.&quot;

if this is the input of MR task what can be the key for the line 4

12. A vacuous case that has no partition, application runs on map tasks alone is

called \_\_\_

13. The number of map tasks that were launched can be determined by using the

function.

14. The number of reduce tasks that were launched can be determined by using the

function.

15. The number of uber tasks that were launched can be determined by using the

function.

16. The number of map tasks that were failed for potential causes can be

determined by using the function.

17. The number of reduce tasks that were failed for potential causes can be

determined by using the function.

18. The number of uber tasks that were failed for potential causes can be

determined by using the function.

19. &quot;Hadoop is a distributed file system

Hadoop integrates with other storage system&quot; is the input then what can be the

value of &#39;file&#39; in reduce phase in a MapReduce word count problem.

20.

&quot;Hadoop is a distributed file system

Hadoop integrates with other storage system&quot; is the input then how many

number reduce tasks will be there for a typical MapReduce word count problem.

21. Pig language used to express data flows called

22. Two environments for executing the programs in Pig are

23. Installation of Pig on a workstation can be done using the

following command

24. The Pig Interactive shell can be started using the following

command

25. Pig commands can run a file in a local mode\_\_\_\_

26. \_\_\_\_ is an Eclipse plug-in that provides an environment for

developing Pig programs.

27. \_\_\_ is a row of data in a database table.

28. Structure of a relation&#39;s schema can be given by the command

29. The contents of a relation can be viewed by the following

command

30. The command used to group the records by the condition fields

Answers:

delegation token

block access token

namenode

list(k2,v2)

list(k3,v3)

hash partitioner

max(minimumSize, min(maximumSize, blockSize))

0

33

57

89

Zero Reducer

TOTAL\_LAUNCHED\_MAPS

TOTAL\_LAUNCHED\_REDUCES

TOTAL\_LAUNCHED\_UBERTASKS

NUM\_FAILED\_MAPS

NUM\_FAILED\_REDUCES

NUM\_FAILED\_UBERTASKS

1

10

Pig latin

local execution in a single JVM and distributed execution on a Hadoop Cluster

%tar xzf pig-x.y.z.taz.gz

Grunt

script.pig

PigPen

Tuple

DESCRIBE

DUMP

GROUP

Revision-II

FBD

1. A \_\_\_ is just an unordered collection of tuples which in Pig Latin

is represented using curly braces

2. If (1950,0,1)

(1950,22,1)

(1950,-11,1)

(1949,111,1)

(1949,78,1) is the input what is the output of grunt&gt; DUMP

max\_temp;

3. The command provides a concise understanding of the query in

a detailed way

4. The command that prints the logical and physical plans in Pig is

5. If the value cannot be cast to the type declared in the schema it

will substitute using a \_\_\_\_ character

6. Which among the following is an EVAL function

7. Which among the following is an FILTER function

8. Which among the following is an LOAD function

9. Which among the following is an EVAL function

10. \_\_\_ provide a way to package reusable pieces of PIG latin code

11. grunt&gt; DUMP A;

(Joe,cherry,2)

(Ali,apple,3)

grunt&gt; B = FOREACH A GENERATE $0, $2+1, &#39;Constant&#39;; then

output of B is

grunt&gt; DUMP B;

12. grunt&gt; DUMP A;

(Joe,cherry,2)

(Ali,apple,3)

grunt&gt; C = STREAM A THROUGH `cut -f 2`; then output of C is

grunt&gt; DUMP C;

13. \_\_\_\_ is implemented by distributing the small input to all the

mappers and performing a map-side join using an in-memory

lookup table against the fragmented relation.

14. The first field of each tuple is the key and the remaining fields

are bags of tuples from the relation with a matching key using

the \_\_\_ command

15. \_\_\_ joins every tuple a first relation with every tuple in a second

relation

16.

grunt&gt; DUMP A;

(2,3)

(1,2)

(2,4)

B = ORDER A BY $0, $1 DESC;

grunt&gt; DUMP B; then output will be

17. grunt&gt; DUMP B;

(1,2)

(2,4)

(2,3)

grunt&gt; D = LIMIT B 2;

grunt&gt; DUMP D; then output will be

18. grunt&gt; DUMP A;

(2,3)

(1,2)

(2,4)

grunt&gt; DUMP B;

(z,x,8)

(w,y,1)

grunt&gt; C = UNION A, B;

grunt&gt; DUMP C; then output will be

19. grunt&gt; DUMP A;

(2,Tie)

(4,Coat)

(3,Hat)

(1,Scarf)

grunt&gt; DUMP B;

(Joe,2)

(Hank,4)

(Ali,0)

(Eve,3)

(Hank,2)

grunt&gt; C = JOIN A BY $0, B BY $1;

grunt&gt; DUMP then output will be

20.

grunt&gt; DUMP A;

(2,Tie)

(4,Coat)

(3,Hat)

(1,Scarf)

grunt&gt; DUMP B;

(Joe,2)

(Hank,4)

(Ali,0)

(Eve,3)

(Hank,2)

grunt&gt; C = JOIN A BY $0 LEFT OUTER, B BY $1;

grunt&gt; DUMP then output will be

21. grunt&gt; DUMP A;

(2,Tie)

(4,Coat)

(3,Hat)

(1,Scarf)

grunt&gt; DUMP B;

(Joe,2)

(Hank,4)

(Ali,0)

(Eve,3)

(Hank,2)

grunt&gt; C = COGROUP A BY $0, B BY $1;

grunt&gt; DUMP then output will be

22.

grunt&gt; DUMP A;

(2,Tie)

(4,Coat)

(3,Hat)

(1,Scarf)

grunt&gt; DUMP B;

(Joe,2)

(Hank,4)

(Ali,0)

(Eve,3)

(Hank,2)

grunt&gt; C = COGROUP A BY $0 INNER, B BY $1;

grunt&gt; DUMP then output will be

23. \_\_\_\_ is a process by which Pig will pipe relations to standard

input of a process and collect the output relation from the

process’s standard output.

24. \_\_\_\_ command removes the leading and trailing whitespaces

from the chararray values in Pig

25.

Among the following choose the drawback of Pig

26. Pig was created by

27. In which year first release of Pig happened

28. Which among the following is not the component of Pig

29. Any single value in Pig Latin, irrespective of their data, type is

known as an \_\_\_\_.

30. A piece of data or a simple atomic value in Pig is known as a

Answers:

BAG

"(1949,111)

(1950,22)"

ILLUSTRATE

EXPLAIN

e'

DIFF

IsEmpty

PigStorage

CONCAT

Macro

"(Joe,3,Constant)

(Ali,4,Constant)"

Cherry apple

fragment replicate join

cogroup

cross

"(1,2)

(2,4)

(2,3)"

"(1,2)

(2,4)"

"(2,3)

(1,2)

(2,4)

(z,x,8)

(w,y,1)"

"(2,Tie,Joe,2)

(2,Tie,Hank,2)

(3,Hat,Eve,3)

(4,Coat,Hank,4)"

"(1,Scarf,,)

(2,Tie,Joe,2)

(2,Tie,Hank,2)

(3,Hat,Eve,3)

(4,Coat,Hank,4)"

"(0,{},{(Ali,0)})

(1,{(1,Scarf)},{})

(2,{(2,Tie)},{(Joe,2),(Hank,2)})

(3,{(3,Hat)},{(Eve,3)})

(4,{(4,Coat)},{(Hank,4)})"

"(1,{(1,Scarf)},{})

(2,{(2,Tie)},{(Joe,2),(Hank,2)}) (3,{(3,Hat)},{(Eve,3)})

(4,{(4,Coat)},{(Hank,4)})"

Streaming

Trim

query optimization

Yahoo!

2008

Profiler

Atom

field

Revision-III FBD

1. Exit the interactive Grunt shell using the command

2. The \_\_\_ operator is used to display the logical, physical, and MapReduce

execution plans of a relation.

3. The \_\_\_ operator is used to group the data in one or more relations. It collects

the data having the same key.

4. \_\_\_\_is used to join a table with itself as if the table were two relations,

temporarily renaming at least one relation.

5.

The \_\_\_ operator of Pig Latin is used to merge the content of two relations.

6. The \_\_\_ operator is used to split a relation into two or more relations.

7. The \_\_\_\_ operator is used to remove redundant (duplicate) tuples from a

relation.

8. The \_\_\_\_ operator is used to generate specified data transformations based on

the column data.

9. The \_\_\_ operator is used to display the contents of a relation in a sorted order

based on one or more fields.

10. To calculate the highest value for a column (numeric values or chararrays) in a

single-column bag.

11. To get the number of elements in a bag, while counting the number of tuples

in a bag.

12. To compute the average of the numerical values within a bag.

13. To concatenate two or more expressions of same type.

14.

To compute the number of elements based on any Pig data type.

15.

To subtract two bags. It takes two bags as inputs and returns a bag which

contains the tuples of the first bag that are not in the second bag.

16.

To get the total of the numeric values of a column in a single-column bag.

17. To split a string (which contains a group of words) in a single tuple and return a

bag which contains the output of the split operation.

18.

To verify whether a given string ends with a particular substring.

19. Accepts two string parameters and verifies whether the first string starts with

the second.

20.

Returns a substring from a given string.

21.

Returns the first occurrence of a character in a string, searching forward from a

start index.

22.

To replace existing characters in a string with new characters.

23.

Returns the date-time object of the current time.

24.

Returns the day of a month from the date-time object.

25.

Returns the hour of a day from the date-time object.

26. Returns the month of a year from the date-time object.

27. To get the absolute value of an expression.

28. This function is used to get the cube root of an expression.

29. This function is used to get the Euler’s number e raised to the power of x.

30. To get the natural logarithm (base e) of an expression.

Answers:

ctrl+d

EXPLAIN

GROUP

Self-Join

Union

Split

Distinct

Foreach

order by

max()

count()

avg()

concat()

size()

substract()

sum()

tokenize()

endswith()

startswith()

substring()

indexof()

replace()

CurrentTime()

GetDay()

GetHour()

GetMonth()

abs()

cbrt()

exp()

log()

Revision 4

FBD

1. Installation of hive on a workstation can be done using the following command

2.

Command to display tables in Hive

3.

The command runs a specified file using the \_\_\_ option

4.

The command used for running the inline commands using the \_\_\_ option

5.

\_\_\_ command displays the current value of any property

6.

The default service an interface in Hive

7.

\_\_\_ runs Hive as a server exposing a Thrift Service for various range of clients

8.

\_\_\_ a convenient way to run java application that includes Hadoop and Hive classes

9.

HWI stands for

10.

Metadata such as table schemas are stored in a database called

11.

\_\_\_ client makes it easy to run hive commands from wide range of programming languages

12.

\_\_\_ will connect to a Hive server running a separate process at the given host and port

13.

\_\_ allows to read data from a table and also write into HDFS in Hive

14.

Hive is not\_\_\_

15.

\_\_\_ means dropping respective tables before dropping the database.

16.

Command to delete a table named &#39;employee&#39;

17. Hive organizes tables into \_\_\_\_ a way of dividing atable into coarse-grained parts based on the value of a

partition column

18. Tables are further subdivided into \_\_\_ to give extra structure to the data that may be used for more

efficient queries

19.

to scan relevant partitions hive performs \_\_\_

20.

CREATE TABLE bucketed\_users (id INT, name STRING)

CLUSTERED BY (id) SORTED BY (id ASC) INTO 4 BUCKETS; in this command &#39;4&#39; indicates

21.

output of the command hive&gt; SELECT round(2.6) from temp;

22.

output of the command hive&gt; SELECT floor(2.6) from temp;

23.

output of the command hive&gt; SELECT ceil(2.6) from temp;

24.

write a command for extracting the salary oa an employee greater than 30000 in hive is

25.

Generate a query to retrieve the employee details in order by using Department name.

26.

Generate a query to retrieve the number of employees in each department.

27.

The tables created in hive are stored as

28.

Which of the following is not a complex data type in Hive?

29.

When a Hive query joins 3 tables, How many mapreduce jobs will be started?

30.

An element in a STRUCT column in hive is referred

**Answers:**

%tar xzf hive-x.y.z.taz.gz

SHOW TABLES;

-f'

-e'

Set

Cli

Hiveserver

Jar

hive web interface

metastore

thrift

JDBC driver

SerDe

OLTP

cascade

DROP TABLE IF EXISTS employee;

Partitions

Buckets

pruning

number of buckets

3

2

3

hive> SELECT \* FROM employee WHERE salary>30000;

hive> SELECT Id, Name, Dept FROM employee ORDER BY DEPT;

hive> SELECT Dept,count(\*) FROM employee GROUP BY DEPT;

a subdirectory under the database directory

matrix

2

Dot

Revision 5

1.HDFS works by breaking large files into smaller pieces called \_\_\_\_

2 . The complete collection of all the files in the cluster is sometimes referred to as the file system \_\_\_\_.C94

3 . Expand SSD

4 . \_\_\_\_ is also responsible for tracking job status and progress within its node.

5. \_\_\_ is a tool that offers the capability to extract data from non-Hadoop data stores

6 . \_\_\_ is a distributed, nonrelational (columnar) database that utilizes HDFS as its persistence store.

7. \_\_\_\_ is managing groups of nodes in service to a single distributed application, it is best

implemented across racks

8. \_\_\_ is designed for scalability and can continually add more resources to a system to handle

extremely large amounts of data in an efficient way.

9. \_\_\_ is designed to work with existing queuing and database technologies.

10. S4 stand for

11. Is the data correct and accurate for the intended usage? Is defined as

12. Are the results meaningful for the given problem space? Is defined as

13. How long do you need to store this data? Is defined as

14. The Namenode receives \_\_\_\_\_ protocol from each DataNode in the cluster.

15. The typical block size is in the Hadoop cluster\_\_\_\_

16. \_\_\_\_\_ file system introduced by Google in 2002 to work for a web crawler

17. Hadoop is coined by \_\_\_\_\_\_

18. Namenodes uses \_\_\_\_ to keep track of data nodes in the Hadoop cluster

19. If NameNode is failed, File System metadata can be recovered from the last saved \_\_\_\_\_ on the Secondary NameNode

20. Apache Spark is used for \_\_\_\_\_

21. Which data was used by the analyst that was provided with the right security and

permissions?

22 .Four V’s of Big Data are:

23. Which of the following is not an Open Source Big Data Tool

24. Computers are best at learning

25. The Cloud Platform of Amazon is \_\_\_\_\_\_\_\_\_\_\_\_\_?

26. What does commodity Hardware in Hadoop world mean?

27. Which of the following are NOT big data problem(s)?

28. What does “Velocity” in Big Data mean?

Answers:

Blocks

name space

solid state disk

node manager

Sqoop

Hbase

Zookeeper

Flume

storm

Simple Scalable Streaming System

Validity

veracity

volatility

Heartbeat

128

Nutch

Dong Cutting

rack id

FSImage

Analytics

meta data

Volume, Velocity, Variety and Veracity

Embedded C Programming

Facts

AWS

Low specifications Industry grade hardware

Both a and b

Speed of storing and processing data