UE18CS322-Big Data- Unit 2

Question Bank and answers

In terms of Ecosystem Tools, do know why the specific tool is required and the possible operations that can be performed using the tool.

PIG:

1. **How does PIG differs from Map Reduce?**

Pig	MapReduce
A dataflow language	A data processing paradigm
High level language and flexible	Low level language and rigid
Performing Join, filter, sorting or ordering operations are quite simple	Relatively difficult to perform Join, filter, sorting or ordering operations between datasets
Programmer with a basic knowledge of SQL can work conveniently	Complex Java implementations require exposure to Java language

Uses multi-query approach, thereby reducing the length of the codes significantly	Require almost 20 times more the number of lines to perform the same task
No need for compilation for execution; operators convert internally into MapReduce jobs	Long compilation process for Jobs
Provides nested data types like tuples, bags and maps	No such data types

2. **Contrast PIG and SQL**

Pig	SQL
Pig Latin is a procedural language	A declarative language
Schema is optional, stores data without assigning a schema	Schema is mandatory
Nested relational data model	Flat relational data model
Provides limited opportunity for Query optimization	More opportunity for query optimization

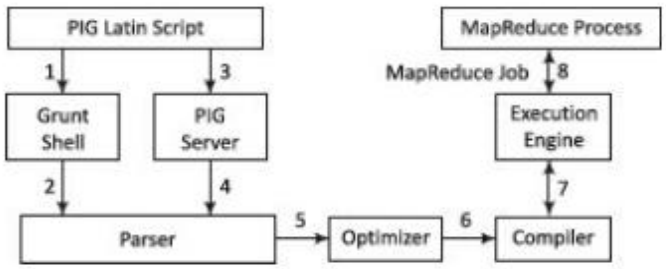
3. **Contrast PIG and HIVE**

Pig	SQL
Pig Latin is a procedural language	A declarative language
Schema is optional, stores data without assigning a schema	Schema is mandatory
Nested relational data model	Flat relational data model
Provides limited opportunity for Query optimization	More opportunity for query optimization

4. What are the different ways to execute a PIG script?

The three ways to execute scripts are:
Grunt Shell: An interactive shell of Pig that executes the scripts.
Script File: Pig commands written in a script file that execute at Pig Server.
Embedded Script: Create UDFs for the functions unavailable in Pig built-in operators. UDF can be in other programming languages. The UDFs can embed in Pig Latin Script file.

5. Explain the PIG architecture.



The Parser performs type checking and checks the script syntax. The output is a Directed Acyclic Graph (DAG).
Optimizer: The DAG is submitted to the logical optimizer. The optimization activities, such as split, merge, transform and reorder operators execute in this phase. The optimization is an automatic feature. The optimizer reduces the amount of data in the pipeline at any instant of time, while processing the extracted data.
Optimization Functions:
Push Up Filter
PushDownForEachFlatten:
ColumnPruner
MapKeyPruner:
LimitOptimizer:

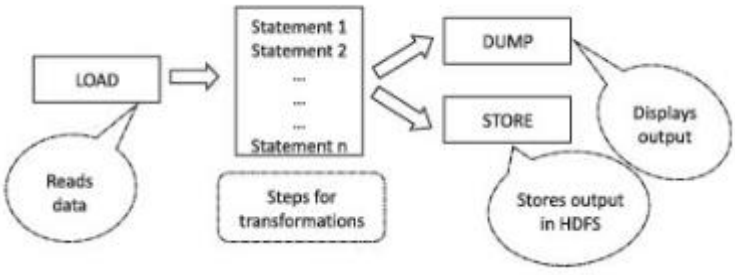
Compiler The compiler compiles after the optimization process. The optimized codes are a series of MapReduce jobs.
Execution Engine Finally, the MapReduce jobs submit for execution to the engine. The MapReduce jobs execute and it outputs the final result.

6. Explain the data models in PIG.

Pig Latin supports primitive data types which are atomic or scalar data types. Atomic data types are int, float, long, double, char [], byte []. The language also defines complex data types. Complex data types are tuple, bag and map.

Data type	Description	Example
bag	Collection of tuples	{(1,1), (2,4)}
tuple	Ordered set of fields	(1,1)
map (data map)	Set of key-value pairs	[Number#1]
int	Signed 32-bit integer	10
long	Signed 64-bit integer	10L or 10l
float	32-bit floating point	22.7F or 22.7f
double	64-bit floating point	3.4 or 3.4e2 or 3.4E2
chararray	Char [], Charac-ter array	data analytics
bytearray	BLOB (Byte array)	ff00

7. Order of Processing in PIG?



8. What are the different type of Operators in PIG?

Operators in Pig Latin

Arithmetic Operators	+	-	*	/	%
Used for	Addition	Subtraction	Multiplication	Division	Remainder

Comparison Operators	==	!=	<	>	≤	≥
Used for	Equality	Not equal	Less than	Greater than	Less than and equal to	Greater than and equal to

Boolean Operators	AND	OR	NOT
Used for	Logical AND	Logical OR	Logical NOT

9. Write a UDF 'IsCorrectAge' which checks if the age given is correct or not. UDF should return a Boolean value: True or False. If the Tuple is null or zero then also it should return False. Use Java. Create a JAR file and then export. Later register the JAR file. The JAR files are in the library files of Apache Pig at the time of loading.

```
public class IsCorrectAge extends FilterFunc {
@Override
    Public Boolean exec (Tuple tuple) throws IOException {
    if (tuple == null || tuple.size() == 0) {
          return false;
    }
    try {
          Object object= tuple.get(0);
          if (object == null) {
                return false;
          }
          Int i = (Integer) object;
          if (i == 18 || i == 20 || i == 21 || i == 25) {
                return true;
          }
          else {
                return false;
          }
    } catch (ExecException e) {
          throw new IOException(e);
    }
    }
}
```

Note: Very important you need to understand all the Group, Join, Filter, Limit, Order by, parallel, sort and split. How it can be performed in PIG with use cases mentioned in the Lecture slides. Given a use case you should be knowing how to choose the relevant operators and operations in PIG that completes your study on PIG.

HIVE:

1. Explain the components of HIVE.

Components of Hive architecture are:

Hive Server (Thrift) – An optional service that allows a remote client to submit requests to Hive and retrieve results. Requests can use a variety of programming languages. Thrift Server exposes a very simple client API to execute HiveQL statements.

Hive CLI (Command Line Interface) – Popular interface to interact with Hive. Hive runs in local mode that uses local storage when running the CLI on a Hadoop cluster instead of HDFS.

Web Interface – Hive can be accessed using a web browser as well. This requires a HWI Server running on some designated code. The URL http:// hadoop:<port no.> / hwi command can be used to access Hive through the web.

Metastore – It is the system catalog. All other components of Hive interact with the Metastore. It stores the schema or metadata of tables, databases, columns in a table, their data types and HDFS mapping. Hive Driver – It manages the life cycle of a HiveQL statement during compilation, optimization and execution.

2. Compare HIVE and RDBMS

Characteristics	Hive	RDBMS
Record level queries	No Update and Delete	Insert, Update and Delete
Transaction support	No	Yes
Latency	Minutes or more	In fractions of a second
Data size	Petabytes	Terabytes
Data per query	Petabytes	Gigabytes
Query language	HiveQL	SQL
Support JDBC/ODBC	Limited	Full

3. Consider a table T with eight-column and four-row table. Partition the table, convert in RC columnar format and serialize.

   Solution

   Firstly, divide the table in four parts, tr1, tr2, tr3 and tr4 horizontally row-wise. Each sub-table has one row and eight columns. Now, convert each sub-table tr1, tr2, tr3 and tr4 into columnar format, or RC File records [Recall Example 3.7 on how RC file saves each row-group data in a format using SERDE (serializer/des-serializer)]. Each sub-table has eight rows and one column. Each column can serially send data one value at an instance. A column has eight key-value pairs with the same key for all the eight.

4. A table toy_tbl contains many values for categories of toys. Assume the number of buckets to be created = 5. Assume a table for Toy_Airplane of product code 10725.

   How will the bucketing enforce?

How will the bucketed table partition toy_airplane_10725 create five buckets?
How will the bucket column load into toy_tbl?
How will the bucket data display?
Solution
#Enforce bucketing set hive.enforce.bucketing=true;
#Create bucketed Table for toy_airplane of product code 10725 and create cluster of 5 buckets CREATE TABLE IF NOT EXISTS toy_airplane_10725(ProductCategory STRING, ProductId INT, ProductName STRING, PrdocutMfgDate YYYY-MM-DD, ProductPrice_US$ FLOAT) CLUSTERED BY (Price) into 5 buckets; # Load data to bucketed table. FROM toy_airplane_10725 INSERT OVERWRITE TABLE toy_tbl SELECT ProductCategory, ProductId, ProductName, PrdocutMfgDate, ProductPrice; To display the contents for Price_US$ selected for the ProductId from the second bucket. SELECT DISTINCT ProductId FROM toy_tbl_buckets TABLE FOR 10725(BUCKET 2 OUT OF 5 ON Price_US$);

5. I am having around 500 tables in a database. I want to import all the tables from the database except the tables named Table 498, Table 323 and Table 199. How can we do this without having to import the tables one by one?
6. I want to see the present working directory in UNIX from Hive. Is it possible to run this command from Hive?
7. What is the use of explode in Hive?
8. Is it possible to change the default location of managed tables in Hive, if so how?
9. Why do we need Hive?
10. What is partitioning? When we may need to customize the default partition? Please explain the scenario with an example.
11. If you run a select * query in Hive, why does it not run MapReduce? Please explain it.
12. What is the difference between external table and managed table?
13. Why do we perform partitioning in Hive? Please explain the advantage of it.
14. Suppose, we create a table that contains details of all the transactions done by the customers of year 2018. CREATE TABLE customer_transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’; Now, after inserting 50,000 tuples in this table, we want to know the total revenue generated for each month. But the problem is, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that we will be taking in order to do so?
15. Explain the data flow in Hive with a diagram. Please describe each and every step.
16. What is the usage of Metastore in Hive. If Metastore is not present in Hive, then what will be the problem?
17. How will you update the rows that are already exported? Write Sqoop command to show all the databases in MySQL server.
18. How to create a table in MySQL and how to insert the values into the table? Please import this table into Hive/HDFS using Apache Sqoop.

Note: Regarding HIVE all the concepts are discussed in the class, for all the questions listed here with no answer, explanation is there in ppts as well as in class videos..do refer to tat for the answers.

For matrix multiplication and Join operations kindly refer to the attached pdf reading chapter 2 which is Ullman book and do refer to the example 2.3 as well.

19. How does a Map task implement using key-value pairs in an input file? What are the uses of Shuffle in processing the aggregates for all the Mapper output by grouping key values of the Mapper output and the value which gets appended in a list of values?
20. How does 'Group By' operate for creating Mapper output? What are the roles of partitioning and combining?
21. How does MapReduce program find the distinct values and count the unique values?
22. How does the MapReduce implement the relational algebraic functions, union, projection, difference, intersection, natural join, grouping and aggregation? Explain each with an example.
23. How do MapReduce tasks implement a matrix multiplication by a vector?

For these questions answers can be found in the pdf as well as in lecture ppts

HBASE and Scoop just know the answers for the below questions

- Consider the following use cases –
    - Data coming from web server logs
    - Web crawling information of web pages and key words
    - Which type of data store will you select for each one and why?

        Ans: for web server logs, we pick HDFS because it is append only and data is not overwritten. For web crawl information, we will pick a database like HBase or Cassandra.
- What is a column family in a columnar database?
- A bank database storing daily bank transactions has chosen to store customers credit information on Cassandra? Is this the right database for storing the data?
- What are regions in HBase database?
- What information is stored in the Metastore of the HBase database?

- Explain what is Sqoop in Hadoop? Please explain the usage
- For each Sqoop copying into HDFS, how many MapReduce jobs and tasks will be submitted? Please explain.
- Explain the significance of using split-by clause in Apache Sqoop.

Note: General guideline to learn module2 would be maximum is discussed in the form of ppts …thourghly if you go through that all sort of questions can be answered.