**Mid Term Test**

**2021F-T1 BDM 1003 - Big Data Tools 01 (DSMM Group 1)**

**Each Question Carries Two Marks: 20**

1. What is Big Data and define 3 V’s?

**Ans:** Big data refers to the data that is large and complex to process using the traditional approach. The 3 V’s in big data are volume velocity and variety. Volume refers to the size of the data, the size of the big data is huge and keeps growing by every minute and hour. Velocity refers to the speed by which the data is collected, with globally connected internet, IOT and other application data is collected in massive speed. Variety refers to the different format and structure in which data is received, the data collected might be relational, non-relational, audio, video and other disparate formats.

1. What is Hadoop and Its Components?

**Ans:**Hadoop is a framework that utilizes the parallel and distributed computing to manage the big data. It is a software that is used to handle big data operation such as storage and analysis (OLTP). The components of Hadoop are

* Hadoop File system (HDFS): It is distributed file system for storage
* Hadoop MapReduce: It is the processing unit or the computation unit for Hadoop
* Hadoop (YARN): Yarn is the resource manager in Hadoop and stands for Yet Another Resource Manager

1. List out the various Hadoop daemons in a Hadoop 1.x

**Ans:** Hadoop 1.x mainly has 3 daemons they are

* Name Node
* Secondary Name Node
* Data Node

1. List out 6 Hadoop Eco-systems?

The components in the Hadoop eco-systems are

* Hive
* Pig
* Sqoop
* Flume
* Scala
* Spark

1. How is HDFS fault tolerant?

HDFS is fault tolerant because of the replication of files among the different machines in the Hadoop Cluster. It creates replica of user’s data on different machine. If any of the machine goes down, user’s data is available from another machine, since it has replicated same data in different machines.

1. How do you define “Rack Awareness” in Hadoop?

Rack Awareness is all about reducing the time to access the data available in nodes. It is the concept that chooses the closer data nodes based on the available rack information. In default configuration Hadoop assumes that all the nodes belong to same rack.

1. What is a Name Node Federation?

The name nodes are independent to each other and don’t require co-ordination among them. For purpose of scaling horizontally, the federation uses multiple name nodes/namespaces. These nodes are federated that means they are independent to each other whereas the data nodes are common to all name nodes and used for common storage.

1. What are the different tables available in Hive

The different table available in Hive are

**Internal Table**

It is default table in Hive and data in these tables are owned by hive. They are also called managed table and can be created without specifying the word external while creating table.

If the managed table or partition is dropped, the table data and the metadata associated with the table is deleted from the HDFS.

**External Table**

Users create the external table when they want to use data outside of hive. External tables are stored outside the warehouse directory and when the table is dropped only the associated meta data is deleted the actual data remains in the same location.

1. Can Hive be used in OLTP systems

Hive cannot be used for OLTP since it does not support insert and update at row level. Hive is generally used for analytical processing so it supports OLAP(online Analytical processing).

1. What is a Hive Metastore

Meta store in Hive is a central repository. It stores the schema and location (meta data for hive table) and partition in relational databases. Client is provided access to meta store using meta store service API.

1. Where does the data of a Hive table get stored

Hive uses managed table to store the data, it stores the data for these table in a sub directory under the directory defined by hive. That is usually (‘/user/hive/warehouse’)

1. Name Pig Running Modes

The two main running modes is pig are

**Local mode**

It is simulation of the map reduce in local machine

**Map reduce Mode**

Map reduce mode is also known as Hadoop mode, where pig is executed in Hadoop cluster. Here pig scripts get converted into sequence of map reduce jobs and get executed in the Hadoop cluster

1. List Simple Data types in Pig

* int: It is signed 32-bit integer.
* long: It is a 64-bit signed integer.
* float: It is a 32-bit floating point.
* double: It is a floating point with double precision.
* chararray: It is character array in unicode UTF-8 format.
* Bytearray: Used to represent bytes.

1. List Complex Data types in Pig

* Tuple: Tuple is an ordered set of fields which may contain different data types for each field.
* Bag: A bag is a collection of a set of tuples and these tuples are a subset of rows or entire rows of a table.
* Map: A map is key-value pairs used to represent data elements.

1. List Schemas in Pig.

The schema in Pig are

With Schema

Without Schema

Partial Schema

Unknown Schema

1. Define BAG, Tuple, Map, Field in Pig

* A bag is a collection of tuples.
* A tuple is an ordered set of fields.
* A Map is the collection or set of key value pair
* A field is a piece of data

1. List Debug Commands in Pig.

**DUMP**: DUMP is used to run (execute) Pig Latin statements and display the results to your screen.

**ILLUSTRATE**: ILLUSTRATE operator is used to review how data is transformed through a sequence of Pig Latin statements. ILLUSTRATE allows you to test your programs on small datasets and get faster turnaround times.

**EXPLAIN**: EXPLAIN operator is used to review the logical, physical, and map reduce execution plans that are used to compute the specified relationship.

**DESCRIBE**: DESCRIBE operator to view the schema of a relation. You can view outer relations as well as relations defined in a nested FOREACH statement

1. Explain Load command with an Example.

The load statement will simply load the data into the specified relation in Pig.

Relation\_name = LOAD 'Input file path' USING function as schema;

Example:

customer = LOAD 'hdfs://localhost:9000/pig\_data/customer\_data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, phone:chararray,

city:chararray );

Here the load command is used to load relation customer from the file customer\_data.txt. Here the tuple is separated using comma specified using pigStorage.

1. Explain Group command with an Example.

The GROUP operator is used to group the data in one or more relations. It collects the data having the same key.

**Syntax:**

Group\_data = GROUP Relation\_name BY age;

**Example:**

Consider a relation customer\_detail

>customer\_detail = LOAD 'hdfs://localhost:9000/pig\_data/customer\_data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:Int, phone:chararray,

city:chararray );

Now to group customer by age we can

>group\_data = GROUP customer\_detail by age;

1. Explain CoGroup command with an Example.

Group and co-group command works in a same fashion and only under lying difference is that the group operator is normally used with one relation, while the cogroup operator is used in statements involving two or more relations.

Example :

Consider we have two relation such as

**customer\_detail** = LOAD 'hdfs://localhost:9000/pig\_data/customer\_data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:Int, phone:chararray,

city:chararray );

**supplier\_detail** = LOAD 'hdfs://localhost:9000/pig\_data/supplier\_data.txt'

USING PigStorage(',')

as ( id:int, firstname:chararray, lastname:chararray, age:Int, phone:chararray,

city:chararray );

Now to co-group customer and supplier by age we can use co-group command

cogroup\_data = COGROUP **customer\_detail** by age, **supplier\_detail** by age;

1. What is Sqoop

Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases

1. What is Sqoop Import?

Sqoop import is command to import data from table to the Hadoop file system as a text or binary file.

**Command:**

sqoop import \

--connect jdbc:mysql://localhost/customerdb \

--username root \

--table customer --m 1

The above command is used to import the customer table from MySQL database server to HDFS.

1. What is the default file format to import data using Apache Sqoop?

The default file format is text file format to import data using Apache Sqoop.

1. How will you list all the columns of a table using Apache Sqoop?

There is no direct way to list all the columns of table in Apache Sqoop, commands like list-column but we can retrieve the columns of the particular table and transform that file to contain the column names of particular table.

1. What is the importance of sqoop eval tool?

Sqoop eval facilitates to run SQL queries in the database. The results are printed in the console.

It helps to see import queries to make sure that the data has been imported as desired.

Sqoop Eval is available for modeling as well as for defining SQL statements.

**Each Question Carries Five Marks: 50**

1. Explain Hadoop 1.x Architecture

Hadoop 1.x has three main deamons. They are

* Name Node
* Secondary Name Node
* Data Node

Diagram

Description automatically generated

Name Node

* + - Name node runs on the master node and there is only one instance of it.
    - Meta data about files that are distributed around the cluster are managed by it.
    - Meta data are read from a file called fsimage and Name node keeps it in memory
    - When the process starts, meta data regarding newly added and removed files are updated in RAM
    - Edit logs are the file in which the periodical changes are written.

Secondary Name Node

* Depending upon the size of the cluster it can either run on master node or a separate node.
* It is not the backup Name Node
* It reads the information written by Name node in edit logs and creates a updated file of a current cluster metadata and sends the file back to Name node so that fs image is updated.Therefore whenever Name Node daemon is restarted it can always find updated information in fsimage file

Data Node

* There are many instances of this process running on various slave nodes(referred as Data nodes)
* It is responsible for storing the individual file blocks on the slave nodes in Hadoop cluster
* Based on the replication factor, a single block is replicated in multiple slave nodes(only if replication factor is > 1) to prevent the data loss
* Whenever required, this process handles the access to a data block by communicating with Name Node
* This process periodically sends heart bits to Name Node to make Name Node aware that slave process is running

**Map reduce (Distributed Computation)**

It uses two daemons job tracker and task tracker to perform map reduce operation. Map reduce operation is carried out in two phase map and reduce. Reduce phase is only started when map phase is completed.

We can see that HDFS has only one Name Node so if that process or machine goes down complete cluster goes down therefore this architecture is considered as single point of failure.

1. Explain Name Node High Availability with QJM

QJM stands for Quorum Journal Manager. Quorum is the generic term used in clustering where we say that cluster is stable. Quorum gives a list of machines and helps to determine the health of the cluster. There are two types of Quorums: Expected Quorum and Calculated Quorum.

In a classic HA cluster, two separate machines are configured as Name Nodes. At any point, one of the Name Nodes will be in Active state and the other will be in a Standby state. The Active Name Node is responsible for all client operations in the cluster, while the Standby is simply acting as a slave, maintaining enough state to provide a fast failover.

The active and standby name node communicate with separate daemons group called journal nodes (JN). The standby node can update and reading the information from the JN. Standby Node keeps all the information available and synchronized with the JNS so that at any point of failure it can start and perform the job that Name Node has been doing.

1. Why do we use HDFS for applications having large data sets and not when there are a lot of small files?

HDFS is efficient for the applications with large datasets rather than application with lots of small files because of the numbers of the meta data required to keep track of each small files.

Large number of small files means large number of metadata related to each files. NameNode keeps the meta data for file system in RAM thus there is limited space available for meta data of files in RAM. In other words, more file will generate more meta data which in turn requires more RAM to store. Ideally it is preferred to have a meta data of a block or a directory only use 150 bytes.

1. Explain Hive Architecture

Hive is A ETL tool built on top of Hadoop to analyze and summarize the large datasets.

The main components of Hive are

1. Hadoop Core Components

Hadoop core components consist of HDFS and map reduce. HDFS is distributed and paraller file system for storage and map reduce is for processing or computation of those stored data.

1. Meta Store

Meta store is the namespace for Hive. All the crucial information related to hive such as table, column, partition’s location are stored in meta store. Meta store is central repository of Hive.

1. Driver

Driver utilizes information present in the meta store to generate the execution plan after parsing the query, performing semantics analysis on them. The execution plan contain the DAG of the stages.

1. Hive Clients

These are the interface through which the queries are submitted to hive. Beeline and hive cli are some of the terminal interfaces. Ambari is one of the web interface.

Diagram

Description automatically generated

1. What applications are supported by Hive

Log Processing: To process the log file from sysytem

Text Mining,: To extract useful info from text file

Document Indexing,: To perform document indexing

Google Analytics,

Sentiment analysis,: Semantic analysis in audio, video and text files.

Predictive Modeling,: to predict the possible outcomes in future

Hypothesis Testing

1. Explain PIG Architecture

The language used to analyze data in Hadoop using Pig is known as pig latin. It is high level data processing language. Pig converts scripts into series of map reduce jobs.

Diagram

Description automatically generated

The major components of pig are

**Parser**

Pig scripts are handled by parser. It performs syntax, type checking and other checks. The output of the parser will be DAG which represents operators and statements.

Optimizer

Logical optimization such as projection and pushdown are carried out by optimizer. The DAG are passed to logical optimizer.

Complier

The complier compiles the logical plan into the series of map reduce jobs

Execution Engine

MapReduce jobs are submitted to Hadoop in a sorted order. Finally, these MapReduce jobs are executed on Hadoop producing the desired results.

1. Differences between Pig and Hive
2. Pig operates on the client side of cluster whereas hive operates on server side of cluster
3. Pig uses pig latin language whereas Hive uses HiveQL language
4. Pig is a procedural data flow language, but Hive is declarative SQl like language
5. Pig is mainly used by researchers and programmers; Hive is mainly used by data analysts.
6. Pig does not support portioning hive supports partitioning
7. What are some of the Apache Pig use cases you can think of

* The web blog can be processed using pig.
* Gathering advertising information from multiple source and analyzing customer behaviour
* Processing health care information
* Stock analysis, to calculate average dividend, trade estimation

1. Explain SQOOP Architecture

Sqoop is mainly used to export and import between relational databases and Hadoop.

Data transfer between Sqoop Hadoop and external storage system is made possible with the help of Sqoop’s connectors.

Sqoop basically have the connectors to connect wide ranges of data vases including MySQL, PostgreSQL, Oracle, SQL Server, and DB2. Each of these connectors knows how to interact with its associated DBMS.

There is also a generic JDBC connector for connecting to any database that supports Java’s JDBC protocol. In addition, Sqoop Big data provides optimized MySQL and PostgreSQL connectors that use database-specific APIs to perform bulk transfers efficiently

Diagram

Description automatically generated

1. What are SQOOP use cases

Sqoop is used to import data from relational database to hadoop and the analysis is carried out on the sqoop.

* Online Marketer uses sqoop to exchange data between Hadoop and IBM Netezza data warehouse application
* Education company apollo group uses sqoop not only to extract data from data base but also to inject data back to database.