Program 1 MapReduce - Counting Records

Question	Answer
What are the roles of the Mapper and Reducer in your program?	Java Mapper emits `<Text("count"), IntWritable(1)>`, Reducer sums values.
What data type does the Mapper output?	Mapper outputs key: Text, value: IntWritable.
What is the key-value pair emitted by the Mapper?	`("count", 1)` for every input line.
What is the role of the job.setJarByClass() method?	Specifies the class with the main method for the job jar.
How do you specify input and output paths in your code?	With `FileInputFormat` and `FileOutputFormat` methods.
What happens if the output directory already exists?	Job fails unless `/output` is deleted.
What does `waitForCompletion(true)` do?	Submits job and waits; `true` enables progress logging.
Can you write this program in Python using Hadoop streaming?	Yes, using Hadoop Streaming with Python scripts.
What does the Reducer do with the values?	Sums up all the `1`s to return total line count.

Question	Answer
What would be the output if your dataset was empty?	Output will still be `count 0`.

Program 2 MapReduce - Maximum Temperature

Question	Answer
How do you extract the year and temperature from each line?	Use `substring()` or regex to extract year and temperature.
How do you handle missing or corrupt temperature values?	Use `if` checks to ignore null or extreme values (e.g., 9999 .
Why use Text and IntWritable instead of String and int?	Hadoop uses Writable types for serialization.
What would be the key and value output from the Mapper?	Key: year, Value: temperature.
How does the Reducer find the max temperature?	Loops through all values, keeps max in `int maxTemp`.
How can you optimize this program for large datasets?	Use Combiner, compress intermediate outputs.

What happens if multiple years have the same max temperature?	Returns one (non-deterministic if same).
Can you modify this to find both max and min temperature?	Add extra logic to track both min and max.
How is Combiner used in such scenarios?	Combiner reduces mapper output size locally.
What is the difference between local and HDFS paths?	Local uses `file:///`, HDFS uses `hdfs:///`.

Program 3 Pig - MovieLens Dataset

Question	Answer
What is the structure of the MovieLens dataset?	Rows with `userId`, `movieId`, `rating`, `timestamp`.
How do you load the data into Pig?	`LOAD 'file' USING PigStorage(',') AS (...);`
What is the difference between GROUP BY and JOIN?	`GROUP BY` aggregates; `JOIN` combines tables.
How do you filter users who rated a movie > 3?	`FILTER data BY rating > 3;`
How is `FOREACH ... GENERATE` used in Pig?	Projects fields: `FOREACH ... GENERATE movieId, COUNT(...);`
How do you calculate average rating for each movie?	`GROUP BY movieId`, then `AVG(rating)` inside `FOREACH`.
What is a relation in Pig?	A dataset (like a table).
What are bags in Pig Latin?	Collections of tuples (grouped data).

Question	Answer
What is the role of DESCRIBE and DUMP?	`DESCRIBE` shows schema; `DUMP` shows output in terminal.
How do you store the output to HDFS?	`STORE result INTO 'output_path' USING PigStorage(',');`

Program 4 Advanced Pig Concepts

Question	Answer
What does grouping by year achieve in Pig?	Allows aggregation per year.
How do you create a bag for grouped data?	Done automatically by `GROUP BY year`.
What function is used to find max or avg temperature?	Use `MAX(temp)` or `AVG(temp)` inside `FOREACH`.
How do you filter for a specific state in Pig?	`FILTER data BY state == 'XY';`
How do you process 3 years’ worth of data?	`FILTER` using `IN` or `OR` for 3 years.
What data type is used for temperature values?	Usually `float` or `int` (dataset-dependent).
What is a `FLATTEN()` in Pig and when is it used?	Removes nesting (useful after grouping).
How do you limit or order Pig results?	Use `LIMIT` and `ORDER BY`.
How do you write a UDF in Pig?	Write Java/Python UDF; register using `REGISTER`.
What’s the difference between DUMP and STORE?	`DUMP` shows immediately; `STORE` saves to HDFS.

Program 5 Hive - MovieLens Facts Extraction

Question	Answer
What are Hive tables: managed vs external?	Managed: Hive controls data; External: Hive stores schema only.
What is CASE statement and how is it used?	`CASE WHEN recommended = 'Y' THEN 1 ELSE 0 END AS recommended_int;`
How do you filter NULL values in Hive?	`WHERE genreid IS NOT NULL;`
How do you select only specific activity types?	`WHERE activity IN (...);`
How do you convert ‘Y’/’N’ to 1/0 in Hive?	`CASE WHEN` to convert `Y/N` to `1/0`.
How can you limit output to 25 rows?	Add `LIMIT 25` at query end.
What is the difference between WHERE and HAVING?	`WHERE` filters pre-grouping; `HAVING` post-grouping.
How do you sort data in Hive?	`ORDER BY column ASC/DESC;`
How is schema-on-read used in Hive?	Schema applied at query time (not storage).
How do you create a partitioned table in Hive?	`PARTITIONED BY (year STRING)` in `CREATE TABLE`.