Question 1: Give an overview of what you understand as the MapReduce technology. Explain how it is similar to the Split-Apply-Combine technology created by Hadley Wickham.

MapReduce is a programming model introduced by Google for processing and generating large data sets on clusters of computers. The MapReduce framework is beneficial because library routines can be used to create parallel programs without any worries about infra-cluster communication, task monitoring or failure handling processes. MapReduce runs on a large cluster of commodity machines and is highly scalable. It has several forms of implementation provided by multiple programming languages, like Java, C# and C++.

The MapReduce framework has two parts:

A function called "Map," which allows different points of the distributed cluster to distribute their work

A function called "Reduce," which is designed to reduce the final form of the clusters' results into one output

The main advantage of the MapReduce framework is its fault tolerance, where periodic reports from each node in the cluster are expected when work is completed.

A task is transferred from one node to another. If the master node notices that a node has been silent for a longer interval than expected, the main node performs the reassignment process to the frozen/delayed task.

The MapReduce framework is inspired by the "Map" and "Reduce" functions used in functional programming. Computational processing occurs on data stored in a file system or within a database, which takes a set of input key values and produces a set of output key values.

Each day, numerous MapReduce programs and MapReduce jobs are executed on Google's clusters. Programs are automatically parallelized and executed on a large cluster of commodity machines. The runtime system deals with partitioning the input data, scheduling the program's execution across a set of machines, machine failure handling and managing required intermachine communication. Programmers without any experience with parallel and distributed systems can easily use the resources of a large distributed system.

MapReduce is used in distributed grep, distributed sort, Web link-graph reversal, Web access log stats, document clustering, machine learning and statistical machine translation.

**The split-apply-combine technology is where you break up a big problem into manageable pieces, operate on each piece independently and then put all the pieces back together. The split-apply-combine technology is similar to the map-reduce strategy for processing large data. In map-reduce, the map step corresponds to split

and apply, and reduce corresponds to combine, although the types of reductions are much richer than those performed for data analysis. Map-reduce is designed for a highly parallel environment, where work is done by hundreds or thousands of independent computers, and for a wider range of data processing needs than just data analysis.

Question 2: Briefly explain three effective method for field research

Field research has a long history. Field research is the collection of raw data outside a laboratory, library, or workplace setting. The approaches and methods used in field research vary across disciplines. Field research is one of the various qualitative methods that market researchers use to better understand customers' needs and wants. Research methods are designed in a detailed, systematic, scientific method for conducting research and obtaining data, or perhaps an ethnographic study utilizing an interpretive framework. Planning the research design is a key step in any Data Science project study.

There are different widely used methods of field research: survey, field research, experiment, and secondary data analysis, or use of existing sources. Every research method comes with plusses and minuses, and the topic of study strongly influences which method or methods are put to use.

Surveys

As a research method, a survey collects data from subjects who respond to a series of questions about behaviours and opinions, often in the form of a questionnaire. The survey is one of the most widely used scientific research methods. The standard survey format allows individuals a level of anonymity in which they can express personal ideas.

Surveys gather different types of information from people. While surveys are not great at capturing the ways people really behave in social situations, they are a great method for discovering how people feel and think—or at least how they say they feel and think. Surveys can track preferences for presidential candidates or reported individual behaviours (such as sleeping, driving, or texting habits) or factual information such as employment status, income, and education levels.

A common instrument is a questionnaire, in which subjects answer a series of questions. For some topics, the researcher might ask yes-or-no or multiple-choice questions, allowing subjects to choose possible responses to each question. This kind of quantitative data—research collected in numerical form that can be counted—are easy to tabulate. Just count up the number of “yes” and “no” responses or correct answers, and chart them into percentages. Questionnaires can also ask more complex questions with more complex answers—beyond “yes,” “no,” or the option next to a checkbox. In those cases, the answers are subjective and vary from person to person.

Case Study

Sometimes a researcher wants to study one specific person or event. A case study is an in-depth analysis of a single event, situation, or individual. To conduct a case study, a researcher examines existing sources like documents and archival records, conducts interviews, engages in direct observation and even participant observation, if possible.

Researchers might use this method to study a single case of, for example, a foster child, drug lord, cancer patient, criminal, or rape victim. However, a major criticism of the case study as a method is that a developed study of a single case, while offering depth on a topic, does not provide enough evidence to form a generalized conclusion. In other words, it is difficult to make universal claims based on just one person, since one person does not verify a pattern. This is why most sociologists do not use case studies as a primary research method.

However, case studies are useful when the single case is unique. In these instances, a single case study can add tremendous knowledge to a certain discipline. For example, a feral child, also called "wild child," is one who grows up isolated from human beings. Feral children grow up without social contact and language, which are elements crucial to a "civilized" child's development. These children mimic the behaviours and movements of animals, and often invent their own language. There are only about one hundred cases of "feral children" in the world.

Experiments

There are two main types of experiments: lab-based experiments and natural or field experiments. In a lab setting, the research can be controlled so that perhaps more data can be recorded in a certain amount of time. In a natural or field-based experiment, the generation of data cannot be controlled but the information might be considered more accurate since it was collected without interference or intervention by the researcher. As a research method, either type of experiment is useful for testing if-then statements: if a particular thing happens, then another particular thing will result. To set up a lab-based experiment, data scientists create artificial situations that allow them to manipulate variables.