Apache Pig Architecture, Hive Working & Infosphere Streams – Detailed Answers

1. Architecture of Apache Pig

Apache Pig is a high-level data flow platform used to analyze large datasets stored in Hadoop. It simplifies MapReduce programming by providing a scripting language known as Pig Latin.

The architecture of Pig consists of multiple components working together to process data efficiently. At the top level, users write Pig Latin scripts, which describe data transformations.

Pig Latin scripts are first processed by the Pig Parser, which checks syntax and performs type checking. After parsing, a Logical Plan is created representing the sequence of operations.

The Logical Optimizer improves performance by applying optimizations such as projection and filter pushdown. The optimized logical plan is then converted into a Physical Plan.

Finally, the physical plan is translated into a series of MapReduce jobs, which are executed on Hadoop. Pig can run in Local Mode or MapReduce Mode depending on the environment.

2. Working of Apache Hive

Apache Hive is a data warehousing tool built on top of Hadoop that provides SQL-like querying using HiveQL. It allows users to query large datasets stored in HDFS without writing MapReduce code.

When a user submits a HiveQL query, it is received by the Hive Driver. The driver manages the query lifecycle and coordinates execution.

The query is sent to the Compiler, which performs syntax and semantic analysis. The Metastore is consulted to fetch metadata such as table schema and location.

The Optimizer improves the execution plan, and the Execution Engine converts the plan into MapReduce, Tez, or Spark jobs.

The results are retrieved from HDFS and returned to the user through the Hive interface.

3. Classification of Apache Pig Operators

Apache Pig operators are used in Pig Latin scripts to process and transform data. They are broadly classified based on their functionality.

- Load and Store Operators – LOAD and STORE are used to read and write data.
- Filtering Operators – FILTER is used to select specific records.
- Grouping Operators – GROUP and COGROUP are used to group data.
- Joining Operators – JOIN is used to combine datasets.
- Sorting Operators – ORDER BY arranges data in sorted order.
- Aggregation Operators – FOREACH and GENERATE are used for aggregation.
- Relational Operators – DISTINCT, LIMIT, and UNION are used for relational processing.

4. Infosphere Streams

IBM Infosphere Streams is a real-time analytics platform designed to process large volumes of streaming data with very low latency.

Unlike traditional batch processing systems, Infosphere Streams processes data in motion. It is widely used in applications such as fraud detection, network monitoring, sensor data processing, and financial analytics.

Infosphere Streams uses a distributed architecture and supports continuous queries. It provides scalability, fault tolerance, and real-time insights.