S. J. P. N. TRUST'S

HIRASUGAR INSTITUTE OF TECHNOLOGY, NIDASOSHI

Accredited at 'A' Grade by NAAC

Programmes Accredited by NBA: CSE, ECE, EEE & ME.

Department of Computer Science & Engineering

Module No.: 4

Lecture No.: 10

Topic: Apache Pig

Course coordinator: Dr. K. B. Manwade

Apache Pig

- Is an abstraction over MapReduce
- Is an execution framework for parallel processing
- Reduces the complexities of writing a MapReduce program
- Is a high-level dataflow language.
- Dataflow language means that a Pig operation node takes the inputs and generates the output for the next node
- Is mostly used in HDFS environment
- Performs data manipulation operations at files at data nodes in Hadoop.

Applications of Apache Pig

- Analyzing large datasets
- Executing tasks involving adhoc processing
- Processing large data sources such as web logs and streaming online data
- Data processing for search platforms. Pig processes different types of data
- Processing time sensitive data loads; data extracts and analyzes quickly.
- For example, analysis of data from twitter to find patterns for user behavior and recommendations.

Features of Apache Pig

- Apache PIG helps programmers write complex data transformations using scripts
- Creates user defined functions (UDFs) to write custom functions which are not available in Pig
- Process any kind of data, structured, semi-structured or unstructured data, coming from various sources
- Reduces the length of codes using multi-query approach
- Handles inconsistent schema in case of unstructured data as well
- Extracts the data, performs operations on that data and dumps the data in the required format in HDFS

Features of Apache Pig

- Performs automatic optimization of tasks before execution.
- Programmers and developers can concentrate on the whole operation without a need to create mapper and reducer tasks separately.
- Reads the input data files from HDFS or the data files from other sources such as local file system, stores the intermediate data and writes back the output in HDFS.
- Pig characteristics are data reading, processing, programming the UDFs in multiple languages and programming multiple queries by fewer codes.
- Pig derives guidance from four philosophies, live anywhere, take anything, domestic and run as if flying.

Differences between Pig and MapReduce

Pig	MapReduce
A dataflow language	A data processing paradigm
High level language and flexible	Low level language and rigid
PerformingJoin, 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	ComplexJava implementations require exposure to Java language
Uses multi-query approach, thereby reducing the length of the codes significant! y	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