

Pig for wrangling Bigdata

Modality: Self-Paced Learning

Duration: 6 Hours

SATV Value:

CLC:

NATU:

SUBSCRIPTION: Learn, Master

Taught by a team which includes 2 Stanford-educated, ex-Googlers and 2 ex-Flipkart Lead Analysts. This team has decades of practical experience in working with large-scale data processing jobs.

Pig is aptly named, it is omnivorous, will consume any data that you throw at it and bring home the bacon!

Let's parse that

omnivorous: Pig works with unstructured data. It has many operations which are very SQL-like but Pig can perform these operations on data sets which have no fixed schema. Pig is great at wrestling data into a form which is clean and can be stored in a data warehouse for reporting and analysis.

bring home the bacon: Pig allows you to transform data in a way that makes it structured, predictable and useful, ready for consumption.

Audience:

- Analysts who want to wrangle large, unstructured data into shape
- Engineers who want to parse and extract useful information from large datasets

Course Objective:

Pig Basics: Scalar and Complex data types (Bags, Maps, Tuples), basic transformations such as Filter, Foreach, Load, Dump, Store, Distinct, Limit, Order by and other built-in functions.

Advanced Data Transformations and Optimizations: The mind-bending Nested Foreach, Joins and their optimizations using "parallel", "merge", "replicated" and other keywords, Co-groups and Semi-joins, debugging using Explain and Illustrate commands

Real-world example: Clean up server logs using Pig

Prerequisites:

Working with Pig requires some basic knowledge of the SQL query language, a brief understanding of the Hadoop eco-system and MapReduce

Course Outline:

You, This Course and Us

Where does Pig fit in?

Pig Basics

Pig Operations And Data Transformations

Advanced Data Transformations

Optimizing Data Transformations

A real-world example

Installing Hadoop in a Local Environment