¿Qué es Apache Pig?

- **Apache Pig** es una plataforma para **procesar y transformar** grandes volúmenes de datos sobre **Hadoop**.
- Usa **Pig Latin**, un lenguaje declarativo que simplifica tareas típicas de ETL (cargar, limpiar, unir, agrupar, agregar).
- **Flujo típico**: *Cargar* datos $\rightarrow$ *Transformar/Filtrar* $\rightarrow$ *Agrupar/Agregar* $\rightarrow$ *Guardar*.
- **Ventajas**: sintaxis simple, menos código que Java MapReduce, ejecuta sobre el clúster (escalable, tolerante a fallos).
- **Cuándo usarlo**: lotes de datos masivos, preparación/limpieza previa a analítica o ML, pipelines repetibles.
- **Limitaciones**: no es para latencias de milisegundos; orientado a *batch* más que a *streaming*.

Ejemplo básico en Pig Latin (promedio por estudiante)

Objetivo: desde `estudiantes.csv` (*nombre*, *materia*, *nota*) calcular el **promedio de notas** por estudiante y guardarlo.

Script Pig Latin

```
-- Cargar datos CSV: nombre, materia, nota
students = LOAD 'estudiantes.csv'
  USING PigStorage(',')
  AS (nombre:chararray, materia:chararray, nota:int);

-- Agrupar por nombre de estudiante
grouped = GROUP students BY nombre;

-- Calcular promedio de nota por estudiante
avg_grade = FOREACH grouped GENERATE
  group AS estudiante,
  AVG(students.nota) AS promedio;

-- Guardar resultados en CSV
STORE avg_grade INTO 'resultados'
  USING PigStorage(',');
```