GDAŃSK UNIVERSITY OF TECHNOLOGY

Introduction to Hadoop – Pig

Wojciech Waloszek

wowal@eti.pg.edu.pl

Department of Software Engineering
Faculty of Electronics, Telecommunication and Informatics
Gdansk University of Technology

GDAŃSK UNIVERSITY OF TECHNOLOGY

Pig

- Pig is a system which allows for executing relational operations on big data,
- Pig commands construct a flow of jobs for Hadoop,
- Within the commands we use aliases, which are relational variables, relational operations, and input-output commands,
- Within the commands we can also use arithmetic operations, comparisons, built-in functions, and user-defined functions (UDF).

GDAŃSK UNIVERSITY OF TECHNOLOGY

Pig working modes

- Pig can work in three modes:
  - Interactive/command line mode, in which the user enter commands,
  - Batch mode, when the user specifies the file containing commands to be run,
  - As a library for the user's application.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Loading Data

- Loading data from a file is scheduled (~done…) by using LOAD function,
- Example: records = LOAD ‘data.txt’;
- With LOAD function the user might specify additional decoding functions, with the use of USING phrase.
- Example: records = LOAD ‘data.txt’ USING PigStorage(‘,’); this command loads comma-separated values,
- There are many other functions like JsonLoader, also UDFs can be used here,
- LOAD also allows for specifying a schema for the data (more about this later…).

GDAŃSK UNIVERSITY OF TECHNOLOGY

Data Types

- Basic types in Pig embrace collections (bag), tuples and simple types,
- LOAD returns a collection of tuples,
- Within a tuple there might be contained both simple-type attributes and complex attributes (bags and tuples),
- Collections are denoted with {}, and tuples with (),
- {(Antek, 1990), (Basia, 1991)} therefore denotes a collection of tuples,
- Tuples may sometimes be „missing” an element, i.e. they might contain an empty value like in (Celina,).

GDAŃSK UNIVERSITY OF TECHNOLOGY

Schemata

- Aliases (relational variables) can be assigned a schema,
- It can be done at the stage of loading the data:
  records = LOAD 'data.txt' AS (name: chararray, year: int);
- The schema of the alias can be shown with use of DESCRIBE, eg.: DESCRIBE records;
- If a schema has not been assigned, we refer to subsequent columns using the symbols $0, $1, $2 …,
- Most of the functions leave the schema (mostly) unchanged, but there are exceptions,
- A schema can also be assigned to a relation that is already loaded with use of FOREACH (more about this… later!).

GDAŃSK UNIVERSITY OF TECHNOLOGY

Types, built-in functions and operations

- Pig handles a large set of simple datatypes, and arithmetic operations and built-in functions that work on those types,
- The examples of simple datatypes are *float*, *int*, *chararray*, *boolean*,
- Built-in function operate on simlpe datatypes and (sometimes) on other Java types,
- Built-in functions can accept and produce text strings, numbers, dates,
- Operators embrace basic arithmetic operations (like +, -, * etc.) and comparisons (like ==, <, >), and additionally *is null* and *is not null*.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Relational operations: FOREACH

- FOREACH *alias* GENERATE *val1*, *val2*,
  eg. recs2 = FOREACH recs GENERATE $0 + $1,
- FOREACH can be used to:
  - projection, by leaving only the selected columns,
  - creating new attributes by using functions and arithmetic operations,
  - assign a schema to a relation, with use of AS:
    recs2 = FOREACH recs GENERATE $0 AS name:chararray, $1 AS year:int

GDAŃSK UNIVERSITY OF TECHNOLOGY

FILTER, LIMIT and ORDER

- FILTER *alias* BY *expression*,
  np. recs2 = FILTER recs BY $1 < 1990,
- FILTER can be use fro tuple selection,
- LIMIT: np. recs2 = LIMIT recs 10
  constrains the relation to the specified number of tuples,
- LIMIT is often use in conjunction with ORDER BY, eg.:
  recs2 = ORDER recs BY $1
- In general every operation „kills” the order, the exception is LIMIT when executed right after ORDER.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Joins

- To perform joins we use JOIN operation,
- Joins are similar to their SQL counterparts,
- Example:
  r1 = {(Antek, 1990), (Basia, 1991), (Celina,1989)}
  r2 = {(1990, Kraków), (1991, Poznań), (1992, Gdańsk)}

  r3 = JOIN r1 BY $1, r2 BY $0;
  r3 = {(Antek,1990,1990,Kraków), (Basia,1991,1991,Poznań)}
- Outer joins can also be performed, but in that case JOIN can be used only on two relational variables, and the type of the join has to be specified in the first part of the command:

  r3 = JOIN r1 BY $1 FULL OUTER, r2 BY $0;

GDAŃSK UNIVERSITY OF TECHNOLOGY

Grouping

- Grouping in Pig works a bit differently from SQL:
  r1 = {(Antek, 1990), (Basia, 1991), (Darek,1990)}
  r2 = GROUP r1 BY $1;
  r2 = {(1990,{(Antek,1990),(Darek,1990)}),
  (1991,{(Basia,1991)})}
- The result of grouping is a relation with the schema (group, *name_of_original_relation*), so in the example (group, r1),
- The first attribute contains values or the grouping expression (BY), the second attribute contains a collection of tuples which have this value of the grouping expression.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Ending the flow

- By formulating subsequent commands we **only** build a plan for execution of Hadoop jobs,
- The situation changes when we use a command for retrieving the results,
- For retrieving the results we use STORE command: STORE *alias* INTO *folder*;
- Just like with LOAD we can add USING phrase to specify the output format (like PigStorage(*separator*)),
- Another (less common, but we will be using it!) command for retrieving the results is DUMP *alias*; this command writes the results to your screen.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Pigs can eat all...

- For many Pig applications UDFs are very important,
- UDFs are usually functions written in Java,
- The functions can be of various types: filters for checking logical conditions, evals for transforming values, and, probably most important, i/o operations (load/store) for loading and storing data,
- Proper use of i/o operations allows for seamless integration of Pig with other big data systems, like Hive (e.g. via HCatalog).