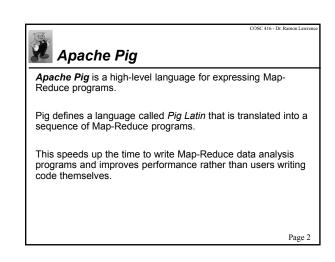
COSC 416 NoSQL Databases

Apache Pig

Dr. Ramon Lawrence University of British Columbia Okanagan ramon.lawrence@ubc.ca



Pig Latin

Pig Latin is very similar to relational algebra.

Each operator takes a relation as input and produces a relation as output. Each statement may use expressions and a schema.

Basic program structure:

- ◆LOAD one or more files from HDFS
- ◆Perform transformation statements
- DUMP (to write output to screen) or STORE to save results to a file. Note that a Map-Reduce program is only generated and run when a DUMP or STORE is encountered.

Page 3

COSC 416 - Dr. Ramon Law

Pig Latin Basic Rules

- 1) Names (aliases) of relations and fields are case-sensitive.
- 2) Function names (e.g. COUNT) are case-sensitive.
- 3) Operator keywords (e.g. LOAD, store, GROUP) are not casesensitive.
- 4) Identifiers must start with a letter and may have digits and underscore.

5) Can reference fields by name or by position. First field is referenced by \$0.

6) A relation is a bag. A bag is a collection of tuples. A tuple is an ordered set of fields. A field is data. Each tuple does not have to have the same fields.

Page 4

COSC 416 - Dr. Ramon La

Pig Latin Operators LOAD

LOAD reads a file from HDFS and references it with a variable. You may specify the loader class and provide a schema to describe the records (both optional).

Syntax:

LOAD 'data' [USING function] [AS schema];

Example:

◆Loads text file using default loader and applies given schema. ◆File is now referenced with identifier **R**.

Page 5

Pig Latin Operators FOREACH (Projection/Iteration)

FOREACH performs column transformations of data such as projections and expression generation.

 Loops through input records one at a time and produces relation of output records.

Syntax:

alias = FOREACH { block | nested_block };

Example:

- X = FOREACH R GENERATE A1, A2; Y = FOREACH R GENERATE A1, SUM(A2), A3+A4;
- ◆Expressions and functions are allowed.

◆Can nest FOREACH to two levels.

◆FLATTEN operator handles nested tuples.

Page 6

Pig Latin Operators FILTER (Selection)

FILTER performs selection (filters) on input.

Syntax:

alias = FILTER alias BY expression; Example: X = FILTER R BY A1 == 3;Y = FILTER R BY A1 > A2;

Pig Latin Operators JÕIN JOIN performs relational inner and outer joins.

Syntax:

alias = JOIN alias BY expression, alias BY expression, ...

Example:

X = JOIN R BY R1, S BY S1;

◆Special settings to handle skew and to select merge joins. ◆May also specify LEFT/RIGHT/FULL OUTER joins.

Page 8

SC 416 - Dr. Ramon La

COSC 416 - Dr. Ramon Lawre

COSC 416 - Dr. Ramon Law

Pig Latin Operators Pig Latin Operators OŘDER BÝ GROUP ORDER BY performs sorting. Svntax: Svntax: alias = ORDER alias BY field [ASC | DESC] Example: Example: = ORDER A BY F1: Sorting is not stable (may change between runs). •Cannot order by fields with complex types or expressions. ◆Can specify * to order by entire tuple. Page 9

COSC 416 - Dr. Ramon Lay

Page 7

GROUP performs relational grouping. alias = GROUP alias BY expression, alias BY expression, ... B = GROUP A BY F1; C = FOREACH B GENERATE group, COUNT(A); C = FOREACH B GENERATE \$0, \$1 May use expressions for grouping. •Can group on multiple relations at the same time. ◆If do not specify a relation, can refer to grouping expression using group or positional notation. Page 10

Pig Latin Operators DUMP/STORE

DUMP writes an output relation to standard output. STORE writes an output relation to a HDFS file.

Svntax:

DUMP alias; STORE alias INTO 'file' [USING function];

Example:

DUMP R; STORE R INTO 'myoutput.txt';

Page 11

COSC 416 - Dr. Ramon Law

Pig Latin Operators Other Useful Operators

DISTINCT removes duplicate tuples in a relation.

SAMPLE partitions a relation into two or more relations. Takes a random sample from the input relation.

 $\ensuremath{\textit{SPLIT}}$ partitions a relation into two or more relations using an expression.

STREAM sends data to an external script or program. UNION computes the union of two or more relations. **REGISTER** registers a JAR that contains UDFs.

Page 12

Pig Latin Operators DESCRIBE and EXPLAIN

DESCRIBE shows the relation for the alias. **EXPLAIN** shows the execution plan. ILLUSTRATE provides an example execution.

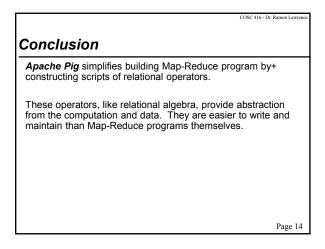
Example: R = LOAD 'myfile.txt' AS (id:int, name:chararray); A = FILTER R BY id > 5; DESCRIBE A; Output: A: {id: int, name: chararray}

EXPLAIN A; Output: Shows an execution plan in Map Reduce.

ILLUSTRATE A; Output: Shows an example output on each stage of the plan.

Page 13

COSC 416 - Dr. Ramon Lawr



COSC 416 - Dr. Ramon Lawrer
Objectives
Understand the basic operators in Pig Latin.
Be able to write queries in Pig to answer English questions.
Page 15