College Sidekick

Home / Computer Science

Pig.ppt

School

Manipal University*

*We are not endorsed by this school

Course

BCD 4

Pages

58

Upload Date

Feb 5, 2024

Uploaded by UltraJay4102 on coursehero.com

Helpful Unhelpful

Introducing Pig

Pig was designed and developed for performing a long series c operations.

The Pig platform is specially designed for handling many ki data, be it structured, semi-structured, or unstructured.

Pig enables users to focus more on what to do than on how to do was developed in 2006 at Yahoo.

It was a part of a research project, the aim of which was to prov easy option to simplify the process of using Hadoo concentrate on examining large datasets instead of wasting ti MapReduce

College Sidekick

- Pig consists of a scripting language, known as Pig Latin, and Latin compiler.
- The scripting language is used to write the code for analyzi data, and the compiler converts the code into the equi MapReduce code.
- So, we can say that Pig automates the process of designing implementing MapReduce applications.
- It becomes easier to write code in Pig compared to programm MapReduce

Pig Architecture

- Pig is simple and easy to use. As stated earlier, Pig is made up components: a scripting language called Pig Latin and the Pi compiler.
- In other words, Pig Latin is the programming language for platform.
- It is a high-level language that is used to write programs f processing and analysis.
- he Pig Latin compiler, on the other hand, is used to convert Latin code into executable code.

College Sidekick

Benefits of Pig

Ease of coding—Pig Latin lets you write complex programs. The c simple and easy to understand and maintain.

It explicitly encodes the complex tasks, involving interrelated data transformations, as data flow sequences.

- **Optimization**—Pig Latin encodes tasks in such a way that they easily optimized for execution.
- This allows users to concentrate on the data processing aspects w bothering about efficiency.
- **Extensibility**—Pig Latin is designed in such a way that it allow create your own custom functions.
- These can be used for performing special tasks. Custom function

College Sidekick

Installing Pig

- You can run Pig from your laptop/desktop computers. It can operate on the machine from which Hadoop jobs are launched can be installed on a UNIX or Windows system.
- Before installing Pig, you need to make sure that your system the following applications:
- Hadoop (version 0.20.2 onwards)
- It can be downloaded from *http://hadoop.apache.org/common/releases.html.*

- The HADOOP_HOME environment variable can b accordingly to indicate the directory where Hadoop is instal
- Java (version 1.6 onwards)
- It can be downloaded from //java.sun.com/javase/downloads/index.jsp.
- The JAVA_HOME environment variable can be set accord to indicate the directory where Java is installed.

College Sidekick

Perform the following steps to download Pig:

1. Download a recent stable release from one of the Apache Down Mirrors.
2. Unpack the downloaded contains the following files:

The Pig script file, Pig, which is located in the bin directory.

The Pig properties file, pig.properties, which is located in the con directory. Add /pig-n.n.n/bin to your path.

Then, use either the export (bash.sh.ksh) or the setenv (tcsh.csh) command, as shown here:

Building a Pig

Repository Perform the following steps to build a Pig repository:

1. Download the Pig code from Apache Subversion (SVN), which available at the following URL: *http://svn.apache.org/repis/asf/pig/trunk*.
2. You can build the code in the working directory. A successfully completed build would result in the creation of the pig.jar file working directory.
3. You can validate the pig.jar file by running a unit test, such as ant test.

College Sidekick

- When you start Pig, the Hadoop properties can be specified with -D option and the Pig properties with the -P option in the Pig interface.
- You can also use the set command to change individual properti in the Grunt mode of Pig.
- The following precedence order is supported by the Pig Latin properties:

- **Local mode**—In this mode, several scripts can run on a sing machine without requiring
- Hadoop MapReduce and Hadoop Distributed File System (HI
- This mode is useful for developing and testing Pig logic.
- The local mode is faster when you are using a small set of da develop or test your code than the MapReduce infrastructure.
- This mode does not require Hadoop.
- The Pig program runs in the context of a local Java Virtual M (JVM).

College Sidekick

- **MapReduce mode**—It is also known as the Hadoop mode.
- In this case, the Pig script gets converted into a series of MapRe jobs, which are then run on the Hadoop cluster.

- The decision whether to use the local mode or the Hadoop r made on the basis of the amount of the data available.
- Suppose you want to perform operations on several terabytes and create a program.
- You may notice that the operations slow down significantl some time.
- The local mode enables you to perform tasks with subsets data in a highly interactive manner.
- You can determine the logic and rectify bugs in your Pig prog

College Sidekick

- Before running the Pig program, it is necessary to know abo pig shell.
- As we all know, without a shell, no one can access the pig built characteristics.
- Pig shell is known as "**Grunt**."
- Grunt is a command shell, which is graphical in nature and for scripting of pig.
- Grunt saves the previously used command in "pig_history" f Home directory.
- There is one handier feature of Grunt: if you are writing a scr grunt, it will automatically complete the keywords that you typing.

Getting Started with Pig Latin

Pig is a high-level programming platform that uses Pig Latin language developing Hadoop MapReduce programs

Pig Latin abstracts its programming from the Java MapReduce idiom extends itself by allowing direct calling of user-defined functions writt Java, Python, JavaScript, Ruby, or Groovy.

Pig Latin is compliant with parallelism, which enables it to handle very l datasets

College Sidekick

Pig translates the Pig Latin script into MapReduce jobs.

The main reasons for developing Pig Latin are as follows:

- Simple—A streamlined method is provided in Pig La interacting with MapReduce.
- This makes the creation of parallel programs simpler for a and processing on the Hadoop cluster.
- Complex tasks may need many interrelated transformations o
- These transformations can be easily encoded as data flow seq in Pig Latin.

- Smart —The Pig Latin compiler transforms a Pig Latin p into a series of Java MapReduce jobs.
- The logic here is to ensure that the compiler can optim execution of these Java MapReduce jobs automatically.
- This allows the user to concentrate on semantics and not on l optimize and access the data.
- Extensible —Pig Latin is extensible, which allows developers functions for addressing their business problems specifically.

College Sidekick

Pig Latin Application Flow

Pig Latin is regarded as a data flow language.

This simply means that we can use Pig Latin to define a data s and a sequence of transformations, which can be applied to th as it moves throughout your application.

In the case of a control flow language, on the other hand, we w sequence of instructions.

We also use concepts such as conditional logic and loops.

College Sidekick

- Pig Latin Structure To explain the Pig Latin structure, let us consider an example of a dataset named Aircraft.
- We want to calculate the total distance covered by aircraft flown different companies.

An explanation for the script in Listing 13.2 is as follows:

- The LOAD operator is used for reading data from HDFS.
- The GROUP operator is used for aggregating input_records.
- The ALL statement is used for aggregating all the tuples into a sin group.
- The FOREACH operator is used for iteration.
- The DUMP statement is used for executing the operators and displaying the results on the screen.

College Sidekick

- Please note the following points about the preceding script:
- The Pig script is much shorter than the MapReduce script for performing a given task
- In Pig, you are not required to be aware of the logic of performi given task.
- The Pig Latin script generally starts with the LOAD operator fo reading data from HDFS.
- Pig has a data model for itself. You need to map the data model files to Pig's data model.

Working with Operators in Pig

- In Pig Latin, relational operators are used for transforming data Different types of transformations include grouping, filtering, sorting, and joining. The following are some basic relational **operators used in Pig:**

1.FOREACH
2.LIMIT
3. ORDER BY
4. ASSERT
5. FILTER
6. DISTINCT
7. JOIN
8. GROUP
9. SPLIT
10. SAMPLE

College Sidekick

FOREACH The FOREACH operator performs iterations over e record to perform a transformation.

When the given expressions are applied, the FOREACH operato generates a new collection of records.

The syntax for using the FOREACH operator is as follows:

College Sidekick

- The preceding example uses the relations 'student', 'rollno', gender.
- The asterisk (*) symbol is used for projecting all the fields.
- You can also use the FOREACH operator for projecting only fields of the student table by using the following script:
- ASSERT The ASSERT operator asserts a condition on the Assertions are used for ensuring that a condition is true on the
- The processing fails if any of the records violate the condition

The syntax for using the ASSERT operator is as follows:

ASSERT alias BY expression [, message]; In the preceding syntax,

alias—Refers to the name of the relation

BY—Refers to a keyword

expression—Refers to a boolean expression

message—Refers to an error message when assertion fails

College Sidekick

FILTER The FILTER operator enables you to use a predicate for selec records that need to be retained in the pipeline.

Only those records will be passed down the pipeline successfully for wh predicate defined in the FILTER statement remains true.

The syntax for using the FILTER operator is as follows:

alias = FILTER alias BY expression;

In the preceding syntax:

alias—Refers to the name of a table

BY—Refers to a keyword

expression—Refers to a boolean expression

College Sidekick

GROUP

Various operators are provided by Pig Latin for group and aggr functions.

The syntax of the GROUP operator in Pig Latin is similar to SQL it is different in functionality when compared to the GROUI clause in SQL.

In Pig Latin, the GROUP operator is used for grouping data in sing multiple relations.

The GROUP BY clause in SQL is used to create a group that ca input directly into single or multiple aggregate functions.

You cannot include a star expression in a GROUP BY column i Latin.

College Sidekick

In the preceding syntax:

alias—Refers to the name of a table

ALL—Refers to the keyword used for inputting all the tuples into group; for example, Z = GROUP A ALL;

BY—Refers to a keyword used for grouping relations by field, tupl expression; for example, X = GROUP A BY f1;

PARTITION BY

partitioner—Describes the Hadoop Partitioner, used for controlling keys that partition intermediate map-outputs.

College Sidekick

- The ORDER BY operator in Pig Latin is used for sorting a give relation, depending on one or more fields.
- The syntax of the ORDER BY operator is as follows:

In the preceding syntax:

- alias—Refers to the name of a relation
- *—Signifies a tuple designator
- field_alias—Refers to a field in the relation
- ASC—Sorts data in ascending order
- DESC—Sorts data in descending order
- PARALLEL n—Enhances the parallelism of a task by mentioni the number of reduce tasks, n Pig supports the ordering on field with simple data types or by using the tuple designator (*). You cannot impose order fields with complex types or by expressions

College Sidekick

- **DISTINCT**
- In Pig Latin, the DISTINCT operator works on the entire recor and not on individual fields. This operator is used for removin duplicate fields from a given set of records.
- The syntax of the DISTINCT operator is as follows:
- ***alias = DISTINCT alias [PARTITION BY partitioner] [PARALLEL n];***
- In the preceding syntax:
- alias—Refers to the name of the relation
- PARTITION BY partitioner— Refers to the Hadoop partitioner
- PARALLEL n—Enhances the parallelism of a task by mentior the number of reduce tasks,

College Sidekick

JOIN

- The JOIN operator in Pig Latin is used for joining two or more relations.
- The joining of two rows is possible in case they have identical If some records in the two rows do not match, these records ca deleted or dropped.
- The following two types of joins can be performed in Pig Latin
- Inner join
- Outer join We will learn about these joins shortly in the chapte

College Sidekick

- BY—Refers to a keyword
- expression—Refers to a field expression
- USING—Refers to a keyword
- replicated—Performs replicated joins
- skewed—Performs skewed joins
- merge—Performs merge joins
- merge-sparse—Performs merge-sparse joins

Outer Join

- When you implement outer joins in Pig, records that do not ha match of the records in the other table are included with null va filled in for the missing fields.
- There are three types of outer joins:
- Left outer join—It returns all the rows from the left table, even there are no matches in the right table.

College Sidekick

The syntax of the LIMIT operator is as follows:

alias = LIMIT alias n;

In the preceding syntax:

- alias—Refers to the name of a relation
- n—Refers to the number of tuples as output, which can be eitl constant, such as 20, or a scalar used in an expression, such as h .add /10

College Sidekick

- It returns the percentage in rows in double values. For example, i the operator returns 0.2, it indicates 20%.
- It is not always likely that the same number of rows will be returned for a particular sample size each time the SAMPLE operator is used.

College Sidekick

In the preceding syntax:

- alias—Refers to the name of a relation
- INTO—Refers to a keyword
- IF—Refers to a keyword
- expression—Refers to an expression
- OTHERWISE—Refers to an optional keyword that designates default relation
- A tuple may be assigned to more than one relation or may not l assigned to any relation, depending on the conditions given in expression.

College Sidekick

operators are as follows:

- Dump —It is used to run statements of Pig Latin and display on screen.
- It is for interactive mode.
- Describe —It is used to view schema of relation.
- Explain —It is used to review logical, physical, and mapredu execution plans, which are used for specified relationship.
- Illustrate —It allows to test your programs on small datasets shows step-by-step execution of sequence of statements.

Turning the Multi-Query Execution

On or Off Multi-query execution remains turned on as by defaul set to "on."

To turn off the multi-query execution and revert to Pig, use these statements "execute-on-dump/store" behavior or use the "-M" no_multiquery" option

From this, the pre-default multi-query execution turns to offline r

College Sidekick

- On working with batch mode execution, the whole pig scri parsed to determine if transitional tasks can be combined to re the overall amount of work that needs to be done;
- parsing is done on priority and then execution starts only parsing is completed (EXPLAIN operator, and the exec and commands are used in this).
- There are two run scenarios, which are optimized and expla below:
- explicit and implicit splits and storing intermediate results.

Explicit and Implicit Splits

College Sidekick

1. This will eliminate the processing of A' multiple times in an enhanced manner.
2. The multiple query execution makes the split non-blocking and a processing to continue.

This helps to reduce the amount of data that has to be stored right split.

3. Multiple query execution allows multiple outputs from a job as c can have one or more queries.

This way some results can be stored as a side effect of the main job is also necessary to make the previous item work.

Working with Functions in Pig

A function can be defined as a set of statements used for per specific tasks.

There are basically two types of functions in Pig: user functions and built-in functions.

As the name suggests, user-defined functions can be created as per their requirements. On the other hand, built-in funct already defined in Pig.

There are mainly five categories of built-in functions in Pig, w as follows:

Eval or Evaluation functions—These functions are used to ev value by using an expression.

College Sidekick

College Sidekick

Load and Store functions —These functions are used to load and extract data. Pig provides a set of built-in load and store functic some of which are described in Table 13.6:

College Sidekick

Return code 1—Used for retrievable errors

Return code 2—All jobs have failed

Return code 3—Some jobs have failed

Uploaded by UltraJay4102 on coursehero.com

College Sidekick

RESOURCES

Study Guides

Study Documents

LEGAL

Copyright, Community Guidelines, DSA & other Legal Resources

Honor Code

Terms

Academic Integrity

College Sidekick

Do not Sell or Share My Personal Info

SUBJECTS

Accounting

Aerospace Engineering

Anatomy

Anthropology

Arts & Humanities

Astronomy

Biology

Business

Chemistry

Civil Engineering

Computer Science

Communications

Economics

Electrical Engineering

English

Finance

Geography

Geology

Health Science

History

Industrial Engineering

Information Systems

Law

Linguistics

Management

Marketing

College Sidekick

Mechanical Engineering

Medicine

Nursing

Philosophy

Physics

Political Science

Psychology

Religion

Sociology

Statistics

SOCIAL

© Learneo, Inc. 2024
*College Sidekick is not sponsored or endorsed by any college or university.