Pig Built-in Functions CHEAT SHEET

IntelliPaat

Built-in Functions

Eval Functions

- AVG (col): Computes the average of the numerical values in a single column of a bag
- CONCAT (string expression1, string expression2): Concatenates two expressions of identical type
- COUNT (DataBag bag): Computes the number of elements in a bag excluding null values
- COUNT STAR (DataBag bag1, DataBag bag 2): Computes the number of elements in a bag including null values.
- DIFF (DataBag bag1, DataBag bag2): It is used to compare two bags, if any element in one bag is not present in the other bag are returned in a bag
- IsEmpty (DataBag bag), IsEmpty(Map map): It is used to check if the bag or map is empty
- Max (col): Computes the maximum of the numeric values or character in a single column bag
- MIN (col): Computes the minimum of the numeric values or character in a single column bag
- DEFINE pluck pluckTuple (expression1): It allows the user to specify a string prefix, and filters the columns which begins with that prefix
- SIZE (expression): Computes the number of elements based on any pig data
- SUBSTRACT (DataBag bag1, DataBag bag2): It returns the bag which does not contain bag1 element in bag2
- SUM: Computes sum of values in one-column bag
- TOKENIZE (String expression, 'field delimiter'): It splits the string and outputs a bag of words

Load and Store Functions

- PigStorage():
  - **PigStorage(field_delimiter)**
  - A = LOAD 'Employee' USING PigStorage('\t') AS (name: chararray, age:int, gpa: float);
  - Loads and stores data as structured text file
- TextLoader():
  - **A = LOAD 'data' USING TextLoader();**
  - Loads unstructured data in UTF 8 format
- BinStorage():
  - **A = LOAD 'data' USING BinStorage();**
  - Loads and stores data in machine readable format
- Handling compression:
  - It loads and stores compressed data in Pig
- JsonLoader, JsonStorage:
  - **A = load 'a.json' using JsonLoader();**
  - It loads and stores JSON data
- Pig dump:
  - **STORE X INTO 'output' USING PigDump();**
  - Stores data in UTF 8 format

Math Functions

- ABS:
  - **ABS(expression)**
  - Returns absolute value of an expression
- COS:
  - **COS(expression)**
  - Returns trigonometric cosine.
- SIN:
  - **SIN (expression)**
  - It returns the sine of an expression.
- CEIL:
  - **CEIL(expression)**
  - Rounds up to the nearest larger integer
- TAN:
  - **TAN(expression)**
  - Returns trigonometric tangent
- ROUND:
  - **ROUND(expression)**
  - Returns value of an expression rounded to an integer (float or long)
- RANDOM:
  - **RANDOM( )**
  - Returns a pseudo random number (type double) >= 0.0 and < 1.0
- Floor:
  - **FLOOR(expression)**
  - Rounds down to the nearest integer.
- CBRT:
  - **CBRT(expression)**
  - It returns the cube root of an expression
- EXP:
  - **EXP(expression)**
  - Returns 'e' raised to the power of 'x'.

String Function

- INDEXOF:
  - **INDEXOF (string, 'character', startIndex)**
  - It returns an index of the first occurrence of a character in a string
- LAST_INDEX:
  - **LAST_INDEX_OF (expression)**
  - It returns an index of the last occurrence of a character in a string
- TRIM:
  - **TRIM(expression)**
  - It returns a copy of the string with leading and trailing whitespaces removed
- SUBSTRING:
  - **SUBSTRING(string, startIndex, stopIndex)**
  - It will return a substring from a given string
- UCFIRST:
  - **UCFIRST(expression)**
  - It will return a string with the first character changed to the upper case
- LOWER:
  - **LOWER(expression)**
  - Converts all characters in a string to lowercase
- UPPER:
  - **UPPER(expression)**
  - Converts all characters in a string to the uppercase

Other Functions

Function	Description
TOTUPLE	Converts expressions to type Tuple
TOBAG	Converts expressions to individual tuples
TOMAP	Converts key expression pairs to Map
TOP	Returns top-n tuples

FURTHERMORE: Hadoop Administrator Online Training