

# LanguageManual UDF

## Hive Operators and User-Defined Functions (UDFs)

- [Hive Operators and User-Defined Functions \(UDFs\)](#)
  - [Built-in Operators](#)
    - [Operators Precedences](#)
    - [Relational Operators](#)
    - [Arithmetic Operators](#)
    - [Logical Operators](#)
    - [String Operators](#)
    - [Complex Type Constructors](#)
    - [Operators on Complex Types](#)
  - [Built-in Functions](#)
    - [Mathematical Functions](#)
      - [Mathematical Functions and Operators for Decimal Datatypes](#)
    - [Collection Functions](#)
    - [Type Conversion Functions](#)
    - [Date Functions](#)
    - [Conditional Functions](#)
    - [String Functions](#)
    - [Data Masking Functions](#)
    - [Misc. Functions](#)
      - [xpath](#)
      - [get\\_json\\_object](#)
  - [Built-in Aggregate Functions \(UDAF\)](#)
  - [Built-in Table-Generating Functions \(UDTF\)](#)
    - [Usage Examples](#)
      - [explode \(array\)](#)
      - [explode \(map\)](#)
      - [posexplode \(array\)](#)
      - [inline \(array of structs\)](#)
      - [stack \(values\)](#)
    - [explode](#)
    - [posexplode](#)
    - [json\\_tuple](#)
    - [parse\\_url\\_tuple](#)
  - [GROUPing and SORTing on f\(column\)](#)
  - [Utility Functions](#)
  - [UDF internals](#)
  - [Creating Custom UDFs](#)



### Case-insensitive

All Hive keywords are case-insensitive, including the names of Hive operators and functions.

In [Beeline](#) or the [CLI](#), use the commands below to show the latest documentation:

```
SHOW FUNCTIONS;
DESCRIBE FUNCTION <function_name>;
DESCRIBE FUNCTION EXTENDED <function_name>;
```



### Bug for expression caching when UDF nested in UDF or function

When [hive.cache.expr.evaluation](#) is set to true (which is the default) a UDF can give incorrect results if it is nested in another UDF or a Hive function. This bug affects releases 0.12.0, 0.13.0, and 0.13.1. Release 0.14.0 fixed the bug ([HIVE-7314](#)).

The problem relates to the UDF's implementation of the `getDisplayString` method, as [discussed](#) in the Hive user mailing list.

## Built-in Operators

### Operators Precedences

| Example             | Operators                                            | Description           |
|---------------------|------------------------------------------------------|-----------------------|
| A[B] , A.identifier | bracket_op( <code>[]</code> ), dot( <code>.</code> ) | element selector, dot |

|                              |                                      |                          |
|------------------------------|--------------------------------------|--------------------------|
| -A                           | unary(+), unary(-), unary(~)         | unary prefix operators   |
| A IS [NOT] (NULL TRUE FALSE) | IS NULL, IS NOT NULL, ...            | unary suffix             |
| A ^ B                        | bitwise xor(^)                       | bitwise xor              |
| A * B                        | star(*), divide(/), mod(%), div(DIV) | multiplicative operators |
| A + B                        | plus(+), minus(-)                    | additive operators       |
| A    B                       | string concatenate(  )               | string concatenate       |
| A & B                        | bitwise and(&)                       | bitwise and              |
| A   B                        | bitwise or( )                        | bitwise or               |

## Relational Operators

The following operators compare the passed operands and generate a TRUE or FALSE value depending on whether the comparison between the operands holds.

| Operator                | Operand types       | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-------------------------|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| A = B                   | All primitive types | TRUE if expression A is equal to expression B otherwise FALSE.                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| A == B                  | All primitive types | Synonym for the = operator.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| A <=> B                 | All primitive types | Returns same result with EQUAL(=) operator for non-null operands, but returns TRUE if both are NULL, FALSE if one of the them is NULL. (As of version <a href="#">0.9.0</a> .)                                                                                                                                                                                                                                                                                                                                  |
| A <> B                  | All primitive types | NULL if A or B is NULL, TRUE if expression A is NOT equal to expression B, otherwise FALSE.                                                                                                                                                                                                                                                                                                                                                                                                                     |
| A != B                  | All primitive types | Synonym for the <> operator.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| A < B                   | All primitive types | NULL if A or B is NULL, TRUE if expression A is less than expression B, otherwise FALSE.                                                                                                                                                                                                                                                                                                                                                                                                                        |
| A <= B                  | All primitive types | NULL if A or B is NULL, TRUE if expression A is less than or equal to expression B, otherwise FALSE.                                                                                                                                                                                                                                                                                                                                                                                                            |
| A > B                   | All primitive types | NULL if A or B is NULL, TRUE if expression A is greater than expression B, otherwise FALSE.                                                                                                                                                                                                                                                                                                                                                                                                                     |
| A >= B                  | All primitive types | NULL if A or B is NULL, TRUE if expression A is greater than or equal to expression B, otherwise FALSE.                                                                                                                                                                                                                                                                                                                                                                                                         |
| A [NOT] BETWEEN B AND C | All primitive types | NULL if A, B or C is NULL, TRUE if A is greater than or equal to B AND A less than or equal to C, otherwise FALSE. This can be inverted by using the NOT keyword. (As of version <a href="#">0.9.0</a> .)                                                                                                                                                                                                                                                                                                       |
| A IS NULL               | All types           | TRUE if expression A evaluates to NULL, otherwise FALSE.                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| A IS NOT NULL           | All types           | FALSE if expression A evaluates to NULL, otherwise TRUE.                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| A IS [NOT] (TRUE FALSE) | Boolean types       | Evaluates to TRUE only if A mets the condition. (since: <a href="#">3.0.0</a> )<br>Note: NULL is UNKNOWN, and because of that (UNKNOWN IS TRUE) and (UNKNOWN IS FALSE) both evaluates to FALSE.                                                                                                                                                                                                                                                                                                                 |
| A [NOT] LIKE B          | strings             | NULL if A or B is NULL, TRUE if string A matches the SQL simple regular expression B, otherwise FALSE. The comparison is done character by character. The _ character in B matches any character in A (similar to . in posix regular expressions) while the % character in B matches an arbitrary number of characters in A (similar to .* in posix regular expressions). For example, 'foobar' like 'foo' evaluates to FALSE whereas 'foobar' like 'foo__' evaluates to TRUE and so does 'foobar' like 'foo%'. |
| A RLIKE B               | strings             | NULL if A or B is NULL, TRUE if any (possibly empty) substring of A matches the Java regular expression B, otherwise FALSE. For example, 'foobar' RLIKE 'foo' evaluates to TRUE and so does 'foobar' RLIKE '^f.*r\$'.                                                                                                                                                                                                                                                                                           |
| A REGEXP B              | strings             | Same as RLIKE.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |

## Arithmetic Operators

The following operators support various common arithmetic operations on the operands. All return number types; if any of the operands are NULL, then the result is also NULL.

| Operator | Operand types    | Description                                                                                                                                                                                                                                                                                                  |
|----------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| A + B    | All number types | Gives the result of adding A and B. The type of the result is the same as the common parent(in the type hierarchy) of the types of the operands. For example since every integer is a float, therefore float is a containing type of integer so the + operator on a float and an int will result in a float. |
| A - B    | All number types | Gives the result of subtracting B from A. The type of the result is the same as the common parent(in the type hierarchy) of the types of the operands.                                                                                                                                                       |
| A * B    | All number types | Gives the result of multiplying A and B. The type of the result is the same as the common parent(in the type hierarchy) of the types of the operands. Note that if the multiplication causing overflow, you will have to cast one of the operators to a type higher in the type hierarchy.                   |
| A / B    | All number types | Gives the result of dividing A by B. The result is a double type in most cases. When A and B are both integers, the result is a double type except when the <a href="#">hive.compat</a> configuration parameter is set to "0.13" or "latest" in which case the result is a decimal type.                     |
| A DIV B  | Integer types    | Gives the integer part resulting from dividing A by B. E.g 17 div 3 results in 5.                                                                                                                                                                                                                            |
| A % B    | All number types | Gives the remainder resulting from dividing A by B. The type of the result is the same as the common parent(in the type hierarchy) of the types of the operands.                                                                                                                                             |
| A & B    | All number types | Gives the result of bitwise AND of A and B. The type of the result is the same as the common parent(in the type hierarchy) of the types of the operands.                                                                                                                                                     |
| A   B    | All number types | Gives the result of bitwise OR of A and B. The type of the result is the same as the common parent(in the type hierarchy) of the types of the operands.                                                                                                                                                      |
| A ^ B    | All number types | Gives the result of bitwise XOR of A and B. The type of the result is the same as the common parent(in the type hierarchy) of the types of the operands.                                                                                                                                                     |
| ~A       | All number types | Gives the result of bitwise NOT of A. The type of the result is the same as the type of A.                                                                                                                                                                                                                   |

## Logical Operators

The following operators provide support for creating logical expressions. All of them return boolean TRUE, FALSE, or NULL depending upon the boolean values of the operands. NULL behaves as an "unknown" flag, so if the result depends on the state of an unknown, the result itself is unknown.

| Operator                   | Operand types | Description                                                                                                                 |
|----------------------------|---------------|-----------------------------------------------------------------------------------------------------------------------------|
| A AND B                    | boolean       | TRUE if both A and B are TRUE, otherwise FALSE. NULL if A or B is NULL.                                                     |
| A OR B                     | boolean       | TRUE if either A or B or both are TRUE, FALSE OR NULL is NULL, otherwise FALSE.                                             |
| NOT A                      | boolean       | TRUE if A is FALSE or NULL if A is NULL. Otherwise FALSE.                                                                   |
| ! A                        | boolean       | Same as NOT A.                                                                                                              |
| A IN (val1, val2, ...)     | boolean       | TRUE if A is equal to any of the values. As of Hive 0.13 <a href="#">subqueries</a> are supported in IN statements.         |
| A NOT IN (val1, val2, ...) | boolean       | TRUE if A is not equal to any of the values. As of Hive 0.13 <a href="#">subqueries</a> are supported in NOT IN statements. |
| [NOT] EXISTS (subquery)    |               | TRUE if the the subquery returns at least one row. Supported as of <a href="#">Hive 0.13</a> .                              |

## String Operators

| Operator | Operand types | Description                                                                                            |
|----------|---------------|--------------------------------------------------------------------------------------------------------|
| A    B   | strings       | Concatenates the operands - shorthand for concat (A, B) . Supported as of <a href="#">Hive 2.2.0</a> . |

## Complex Type Constructors

The following functions construct instances of complex types.

| Constructor Function | Operands                          | Description                                                                                  |
|----------------------|-----------------------------------|----------------------------------------------------------------------------------------------|
| map                  | (key1, value1, key2, value2, ...) | Creates a map with the given key/value pairs.                                                |
| struct               | (val1, val2, val3, ...)           | Creates a struct with the given field values. Struct field names will be col1, col2, ....    |
| named_struct         | (name1, val1, name2, val2, ...)   | Creates a struct with the given field names and values. (As of Hive <a href="#">0.8.0</a> .) |
| array                | (val1, val2, ...)                 | Creates an array with the given elements.                                                    |
| create_union         | (tag, val1, val2, ...)            | Creates a union type with the value that is being pointed to by the tag parameter.           |

## Operators on Complex Types

The following operators provide mechanisms to access elements in Complex Types.

| Operator | Operand types                       | Description                                                                                                                                                                       |
|----------|-------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| A[n]     | A is an Array and n is an int       | Returns the nth element in the array A. The first element has index 0. For example, if A is an array comprising of ['foo', 'bar'] then A[0] returns 'foo' and A[1] returns 'bar'. |
| M[key]   | M is a Map<K, V> and key has type K | Returns the value corresponding to the key in the map. For example, if M is a map comprising of {'f' -> 'foo', 'b' -> 'bar', 'all' -> 'foobar'} then M['all'] returns 'foobar'.   |
| S.x      | S is a struct                       | Returns the x field of S. For example for the struct foobar {int foo, int bar}, foobar.foo returns the integer stored in the foo field of the struct.                             |

## Built-in Functions

### Mathematical Functions

The following built-in mathematical functions are supported in Hive; most return NULL when the argument(s) are NULL:

| Return Type | Name (Signature)                                               | Description                                                                                                                                                                                                                       |
|-------------|----------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DOUBLE      | round(DOUBLE a)                                                | Returns the rounded BIGINT value of a.                                                                                                                                                                                            |
| DOUBLE      | round(DOUBLE a, INT d)                                         | Returns a rounded to d decimal places.                                                                                                                                                                                            |
| DOUBLE      | bround(DOUBLE a)                                               | Returns the rounded BIGINT value of a using HALF_EVEN rounding mode (as of <a href="#">Hive 1.3.0</a> , <a href="#">2.0.0</a> ). Also known as Gaussian rounding or bankers' rounding. Example: bround(2.5) = 2, bround(3.5) = 4. |
| DOUBLE      | bround(DOUBLE a, INT d)                                        | Returns a rounded to d decimal places using HALF_EVEN rounding mode (as of <a href="#">Hive 1.3.0</a> , <a href="#">2.0.0</a> ). Example: bround(8.25, 1) = 8.2, bround(8.35, 1) = 8.4.                                           |
| BIGINT      | floor(DOUBLE a)                                                | Returns the maximum BIGINT value that is equal to or less than a.                                                                                                                                                                 |
| BIGINT      | ceil(DOUBLE a), ceiling(DOUBLE a)                              | Returns the minimum BIGINT value that is equal to or greater than a.                                                                                                                                                              |
| DOUBLE      | rand(), rand(INT seed)                                         | Returns a random number (that changes from row to row) that is distributed uniformly from 0 to 1. Specifying the seed will make sure the generated random number sequence is deterministic.                                       |
| DOUBLE      | exp(DOUBLE a), exp(DECIMAL a)                                  | Returns e <sup>a</sup> where e is the base of the natural logarithm. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                       |
| DOUBLE      | ln(DOUBLE a), ln(DECIMAL a)                                    | Returns the natural logarithm of the argument a. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                           |
| DOUBLE      | log10(DOUBLE a), log10(DECIMAL a)                              | Returns the base-10 logarithm of the argument a. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                           |
| DOUBLE      | log2(DOUBLE a), log2(DECIMAL a)                                | Returns the base-2 logarithm of the argument a. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                            |
| DOUBLE      | log(DOUBLE base, DOUBLE a)<br><br>log(DECIMAL base, DECIMAL a) | Returns the base-base logarithm of the argument a. Decimal versions added in <a href="#">Hive 0.13.0</a> .                                                                                                                        |

|               |                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|---------------|-----------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DOUBLE        | <code>pow(DOUBLE a, DOUBLE p)</code> , <code>power(DOUBLE a, DOUBLE p)</code>                                         | Returns $a^p$ .                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| DOUBLE        | <code>sqrt(DOUBLE a)</code> , <code>sqrt(DECIMAL a)</code>                                                            | Returns the square root of a. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                                                             |
| STRING        | <code>bin(BIGINT a)</code>                                                                                            | Returns the number in binary format (see <a href="http://dev.mysql.com/doc/refman/5.0/en/string-functions.html#function_bin">http://dev.mysql.com/doc/refman/5.0/en/string-functions.html#function_bin</a> ).                                                                                                                                                                                                                                    |
| STRING        | <code>hex(BIGINT a)</code> , <code>hex(STRING a)</code> , <code>hex(BINARY a)</code>                                  | If the argument is an INT or binary, hex returns the number as a STRING in hexadecimal format. Otherwise if the number is a STRING, it converts each character into its hexadecimal representation and returns the resulting STRING. (See <a href="http://dev.mysql.com/doc/refman/5.0/en/string-functions.html#function_hex">http://dev.mysql.com/doc/refman/5.0/en/string-functions.html#function_hex</a> , BINARY version as of Hive 0.12.0.) |
| BINARY        | <code>unhex(STRING a)</code>                                                                                          | Inverse of hex. Interprets each pair of characters as a hexadecimal number and converts to the byte representation of the number. (BINARY version as of Hive 0.12.0, used to return a string.)                                                                                                                                                                                                                                                   |
| STRING        | <code>conv(BIGINT num, INT from_base, INT to_base)</code> , <code>conv(STRING num, INT from_base, INT to_base)</code> | Converts a number from a given base to another (see <a href="http://dev.mysql.com/doc/refman/5.0/en/mathematical-functions.html#function_conv">http://dev.mysql.com/doc/refman/5.0/en/mathematical-functions.html#function_conv</a> ).                                                                                                                                                                                                           |
| DOUBLE        | <code>abs(DOUBLE a)</code>                                                                                            | Returns the absolute value.                                                                                                                                                                                                                                                                                                                                                                                                                      |
| INT or DOUBLE | <code>pmod(INT a, INT b)</code> , <code>pmod(DOUBLE a, DOUBLE b)</code>                                               | Returns the positive value of $a \bmod b$ .                                                                                                                                                                                                                                                                                                                                                                                                      |
| DOUBLE        | <code>sin(DOUBLE a)</code> , <code>sin(DECIMAL a)</code>                                                              | Returns the sine of a (a is in radians). Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                                                  |
| DOUBLE        | <code>asin(DOUBLE a)</code> , <code>asin(DECIMAL a)</code>                                                            | Returns the arc sin of a if $-1 \leq a \leq 1$ or NULL otherwise. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                         |
| DOUBLE        | <code>cos(DOUBLE a)</code> , <code>cos(DECIMAL a)</code>                                                              | Returns the cosine of a (a is in radians). Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                                                |
| DOUBLE        | <code>acos(DOUBLE a)</code> , <code>acos(DECIMAL a)</code>                                                            | Returns the arccosine of a if $-1 \leq a \leq 1$ or NULL otherwise. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                       |
| DOUBLE        | <code>tan(DOUBLE a)</code> , <code>tan(DECIMAL a)</code>                                                              | Returns the tangent of a (a is in radians). Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                                               |
| DOUBLE        | <code>atan(DOUBLE a)</code> , <code>atan(DECIMAL a)</code>                                                            | Returns the arctangent of a. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                                                              |
| DOUBLE        | <code>degrees(DOUBLE a)</code> , <code>degrees(DECIMAL a)</code>                                                      | Converts value of a from radians to degrees. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                                              |
| DOUBLE        | <code>radians(DOUBLE a)</code> , <code>radians(DECIMAL a)</code>                                                      | Converts value of a from degrees to radians. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                                                                                                                              |
| INT or DOUBLE | <code>positive(INT a)</code> , <code>positive(DOUBLE a)</code>                                                        | Returns a.                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| INT or DOUBLE | <code>negative(INT a)</code> , <code>negative(DOUBLE a)</code>                                                        | Returns -a.                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| DOUBLE or INT | <code>sign(DOUBLE a)</code> , <code>sign(DECIMAL a)</code>                                                            | Returns the sign of a as '1.0' (if a is positive) or '-1.0' (if a is negative), '0.0' otherwise. The decimal version returns INT instead of DOUBLE. Decimal version added in <a href="#">Hive 0.13.0</a> .                                                                                                                                                                                                                                       |
| DOUBLE        | <code>e()</code>                                                                                                      | Returns the value of e.                                                                                                                                                                                                                                                                                                                                                                                                                          |
| DOUBLE        | <code>pi()</code>                                                                                                     | Returns the value of pi.                                                                                                                                                                                                                                                                                                                                                                                                                         |
| BIGINT        | <code>factorial(INT a)</code>                                                                                         | Returns the factorial of a (as of Hive 1.2.0). Valid a is [0..20].                                                                                                                                                                                                                                                                                                                                                                               |
| DOUBLE        | <code>cbrt(DOUBLE a)</code>                                                                                           | Returns the cube root of a double value (as of Hive 1.2.0).                                                                                                                                                                                                                                                                                                                                                                                      |
| INT<br>BIGINT | <code>shiftleft(TINYINT SMALLINT INT a, INT b)</code><br><br><code>shiftleft(BIGINT a, INT b)</code>                  | Bitwise left shift (as of Hive 1.2.0). Shifts a b positions to the left.<br>Returns int for tinyint, smallint and int a. Returns bigint for bigint a.                                                                                                                                                                                                                                                                                            |

|               |                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|---------------|------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| INT<br>BIGINT | shiftright<br>(TINYINT SMALLINT INT a,<br>INT b)<br><br>shiftright(BIGINT a, INT b)                  | Bitwise right shift (as of Hive <a href="#">1.2.0</a> ). Shifts a b positions to the right.<br><br>Returns int for tinyint, smallint and int a. Returns bigint for bigint a.                                                                                                                                                                                                                                                                    |
| INT<br>BIGINT | shiftrightunsigned(TINYINT <br>SMALLINT INT a, INT b),<br><br>shiftrightunsigned(BIGINT a,<br>INT b) | Bitwise unsigned right shift (as of Hive <a href="#">1.2.0</a> ). Shifts a b positions to the right.<br><br>Returns int for tinyint, smallint and int a. Returns bigint for bigint a.                                                                                                                                                                                                                                                           |
| T             | greatest(T v1, T v2, ...)                                                                            | Returns the greatest value of the list of values (as of Hive <a href="#">1.1.0</a> ). Fixed to return NULL when one or more arguments are NULL, and strict type restriction relaxed, consistent with ">" operator (as of Hive <a href="#">2.0.0</a> ).                                                                                                                                                                                          |
| T             | least(T v1, T v2, ...)                                                                               | Returns the least value of the list of values (as of Hive <a href="#">1.1.0</a> ). Fixed to return NULL when one or more arguments are NULL, and strict type restriction relaxed, consistent with "<" operator (as of Hive <a href="#">2.0.0</a> ).                                                                                                                                                                                             |
| INT           | width_bucket(NUMERIC<br>expr, NUMERIC min_value,<br>NUMERIC max_value, INT<br>num_buckets)           | Returns an integer between 0 and num_buckets+1 by mapping expr into the ith equally sized bucket. Buckets are made by dividing [min_value, max_value] into equally sized regions. If expr < min_value, return 1, if expr > max_value return num_buckets+1. See <a href="https://docs.oracle.com/cd/B19306_01/server.102/b14200/functions214.htm">https://docs.oracle.com/cd/B19306_01/server.102/b14200/functions214.htm</a> (as of Hive 3.0.0) |

## Mathematical Functions and Operators for Decimal Datatypes



### Version

The decimal datatype was introduced in Hive 0.11.0 ([HIVE-2693](#)).

All regular arithmetic operators (such as +, -, \*, /) and relevant mathematical UDFs (Floor, Ceil, Round, and many more) have been updated to handle decimal types. For a list of supported UDFs, see [Mathematical UDFs](#) in [Hive Data Types](#).

## Collection Functions

The following built-in collection functions are supported in Hive:

| Return Type | Name(Signature)                    | Description                                                                                                                                             |
|-------------|------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| int         | size(Map<K,V>)                     | Returns the number of elements in the map type.                                                                                                         |
| int         | size(Array<T>)                     | Returns the number of elements in the array type.                                                                                                       |
| array<K>    | map_keys(Map<K,V>)                 | Returns an unordered array containing the keys of the input map.                                                                                        |
| array<V>    | map_values(Map<K,V>)               | Returns an unordered array containing the values of the input map.                                                                                      |
| boolean     | array_contains(Array<T>,<br>value) | Returns TRUE if the array contains value.                                                                                                               |
| array<t>    | sort_array(Array<T>)               | Sorts the input array in ascending order according to the natural ordering of the array elements and returns it (as of version <a href="#">0.9.0</a> ). |

## Type Conversion Functions

The following type conversion functions are supported in Hive:

| Return Type                                  | Name (Signature)          | Description                                                                                                                                                                                                                                                                   |
|----------------------------------------------|---------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| binary                                       | binary<br>(string binary) | Casts the parameter into a binary.                                                                                                                                                                                                                                            |
| <b>Expected<br/>"=" to follow<br/>"type"</b> | cast(expr as<br><type>)   | Converts the results of the expression expr to <type>. For example, cast('1' as BIGINT) will convert the string '1' to its integral representation. A null is returned if the conversion does not succeed. If cast(expr as boolean) Hive returns true for a non-empty string. |

## Date Functions

The following built-in date functions are supported in Hive:

| Return Type                         | Name (Signature)                                     | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------------------------------------|------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| string                              | from_unixtime<br>(bigint unixtime[, string pattern]) | <p>Converts a number of seconds since epoch (1970-01-01 00:00:00 UTC) to a string representing the timestamp of that moment in the current time zone(using config "hive.local.time.zone") using the specified pattern. If the pattern is missing the default is used ('uuuu-MM-dd HH:mm:ss' or yyyy-MM-dd HH:mm:ss'). Example: from_unixtime(0)=1970-01-01 00:00:00 (hive.local.time.zone=Etc/GMT)</p> <p>As of Hive 4.0.0 (<a href="#">HIVE-25576</a>), the "hive.datetime.formatter" property can be used to control the underlying formatter implementation, and as a consequence the accepted patterns and their behavior. Prior versions always used <a href="https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html">https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html</a> as the underlying formatter.</p>                              |
| bigint                              | unix_timestamp()                                     | Gets current Unix timestamp in seconds. This function is not deterministic and its value is not fixed for the scope of a query execution, therefore prevents proper optimization of queries - this has been deprecated since 2.0 in favour of CURRENT_TIMESTAMP constant.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| bigint                              | unix_timestamp(string date)                          | <p>Converts a datetime string to unix time (seconds since epoch) using the default pattern(s). The default accepted patterns depend on the underlying formatter implementation. The datetime string does not contain a timezone so the conversion uses the local time zone as specified by "hive.local.time.zone" property. Returns null when the conversion fails. Example: unix_timestamp('2009-03-20 11:30:01') = 1237573801</p> <p>As of Hive 4.0.0 (<a href="#">HIVE-25576</a>), the "hive.datetime.formatter" property can be used to control the underlying formatter implementation, and as a consequence the accepted patterns and their behavior. Prior versions always used <a href="https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html">https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html</a> as the underlying formatter.</p> |
| bigint                              | unix_timestamp(string date, string pattern)          | <p>Converts a datetime string to unix time (seconds since epoch) using the specified pattern. The accepted patterns and their behavior depend on the underlying formatter implementation. Returns null when the conversion fails. Example: unix_timestamp('2009-03-20', 'uuuu-MM-dd') = 1237532400</p> <p>As of Hive 4.0.0 (<a href="#">HIVE-25576</a>), the "hive.datetime.formatter" property can be used to control the underlying formatter implementation, and as a consequence the accepted patterns and their behavior. Prior versions always used <a href="https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html">https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html</a> as the underlying formatter.</p>                                                                                                                              |
| pre 2.1.0: string<br>2.1.0 on: date | to_date<br>(string timestamp)                        | <p>Returns the date part of a timestamp string (pre-Hive 2.1.0): to_date("1970-01-01 00:00:00") = "1970-01-01". As of Hive 2.1.0, returns a date object.</p> <p>Prior to Hive 2.1.0 (<a href="#">HIVE-13248</a>) the return type was a String because no Date type existed when the method was created.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| int                                 | year(string date)                                    | Returns the year part of a date or a timestamp string: year("1970-01-01 00:00:00") = 1970, year("1970-01-01") = 1970.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| int                                 | quarter(date /timestamp /string)                     | Returns the quarter of the year for a date, timestamp, or string in the range 1 to 4 (as of Hive <a href="#">1.3.0</a> ). Example: quarter('2015-04-08') = 2.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| int                                 | month(string date)                                   | Returns the month part of a date or a timestamp string: month("1970-11-01 00:00:00") = 11, month("1970-11-01") = 11.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| int                                 | day(string date)<br>dayofmonth<br>(date)             | Returns the day part of a date or a timestamp string: day("1970-11-01 00:00:00") = 1, day("1970-11-01") = 1.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| int                                 | hour(string date)                                    | Returns the hour of the timestamp: hour('2009-07-30 12:58:59') = 12, hour('12:58:59') = 12.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| int                                 | minute(string date)                                  | Returns the minute of the timestamp.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| int                                 | second<br>(string date)                              | Returns the second of the timestamp.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| int                                 | weekofyear<br>(string date)                          | Returns the week number of a timestamp string: weekofyear("1970-11-01 00:00:00") = 44, weekofyear("1970-11-01") = 44.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| int                                 | extract(field FROM source)                           | <p>Retrieve fields such as days or hours from source (as of Hive <a href="#">2.2.0</a>). Source must be a date, timestamp, interval or a string that can be converted into either a date or timestamp. Supported fields include: day, dayofweek, hour, minute, month, quarter, second, week and year.</p> <p>Examples:</p> <ol style="list-style-type: none"> <li>1. select extract(month from "2016-10-20") results in 10.</li> <li>2. select extract(hour from "2016-10-20 05:06:07") results in 5.</li> <li>3. select extract(dayofweek from "2016-10-20 05:06:07") results in 5.</li> <li>4. select extract(month from interval '1-3' year to month) results in 3.</li> <li>5. select extract(minute from interval '3 12:20:30' day to second) results in 20.</li> </ol>                                                                                                              |

|                                               |                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|-----------------------------------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| int                                           | datediff<br>(string<br>enddate,<br>string<br>startdate)                           | Returns the number of days from startdate to enddate: datediff('2009-03-01', '2009-02-27') = 2.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| pre 2.1.0:<br>string<br><br>2.1.0 on:<br>date | date_add(date/timestamp<br>/string startdate, tinyint<br>/smallint/int<br>days)   | Adds a number of days to startdate: date_add('2008-12-31', 1) = '2009-01-01'.<br><br>Prior to Hive 2.1.0 ( <a href="#">HIVE-13248</a> ) the return type was a String because no Date type existed when the method was created.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| pre 2.1.0:<br>string<br><br>2.1.0 on:<br>date | date_sub(date/timestamp<br>/string startdate, tinyint<br>/smallint/int<br>days)   | Subtracts a number of days to startdate: date_sub('2008-12-31', 1) = '2008-12-30'.<br><br>Prior to Hive 2.1.0 ( <a href="#">HIVE-13248</a> ) the return type was a String because no Date type existed when the method was created.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| timestamp                                     | from_utc_timestamp({any<br>primitive type} ts, string<br>timezone)                | Converts a timestamp* in UTC to a given timezone (as of Hive <a href="#">0.8.0</a> ).<br><br>* timestamp is a primitive type, including timestamp/date, tinyint/smallint/int/bigint, float/double and decimal.<br><br>Fractional values are considered as seconds. Integer values are considered as milliseconds. For example, from_utc_timestamp(2592000.0,'PST'), from_utc_timestamp(2592000000,'PST') and from_utc_timestamp(timestamp '1970-01-30 16:00:00','PST') all return the timestamp 1970-01-30 08:00:00.                                                                                                                                                                                                                                                                                                                                                                                                                           |
| timestamp                                     | to_utc_timestamp({any<br>primitive type} ts, string<br>timezone)                  | Converts a timestamp* in a given timezone to UTC (as of Hive <a href="#">0.8.0</a> ).<br><br>* timestamp is a primitive type, including timestamp/date, tinyint/smallint/int/bigint, float/double and decimal.<br><br>Fractional values are considered as seconds. Integer values are considered as milliseconds. For example, to_utc_timestamp(2592000.0,'PST'), to_utc_timestamp(2592000000,'PST') and to_utc_timestamp(timestamp '1970-01-30 16:00:00','PST') all return the timestamp 1970-01-31 00:00:00.                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| date                                          | current_date                                                                      | Returns the current date at the start of query evaluation (as of Hive <a href="#">1.2.0</a> ). All calls of current_date within the same query return the same value.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| timestamp                                     | current_timestamp                                                                 | Returns the current timestamp at the start of query evaluation (as of Hive <a href="#">1.2.0</a> ). All calls of current_timestamp within the same query return the same value.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| string                                        | add_months<br>(string<br>start_date,<br>int<br>num_months,<br>output_date_format) | Returns the date that is num_months after start_date (as of Hive <a href="#">1.1.0</a> ). start_date is a string, date or timestamp. num_months is an integer. If start_date is the last day of the month or if the resulting month has fewer days than the day component of start_date, then the result is the last day of the resulting month. Otherwise, the result has the same day component as start_date. The default output format is 'yyyy-MM-dd'.<br><br>Before Hive 4.0.0, the time part of the date is ignored.<br><br>As of Hive <a href="#">4.0.0</a> , add_months supports an optional argument output_date_format, which accepts a String that represents a valid date format for the output. This allows to retain the time format in the output.<br><br>For example :<br><br>add_months('2009-08-31', 1) returns '2009-09-30'.<br>add_months('2017-12-31 14:15:16', 2, 'YYYY-MM-dd HH:mm:ss') returns '2018-02-28 14:15:16'. |
| string                                        | last_day<br>(string date)                                                         | Returns the last day of the month which the date belongs to (as of Hive <a href="#">1.1.0</a> ). date is a string in the format 'yyyy-MM-dd HH:mm:ss' or 'yyyy-MM-dd'. The time part of date is ignored.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| string                                        | next_day<br>(string<br>start_date,<br>string<br>day_of_week)                      | Returns the first date which is later than start_date and named as day_of_week (as of Hive <a href="#">1.2.0</a> ). start_date is a string /date/timestamp. day_of_week is 2 letters, 3 letters or full name of the day of the week (e.g. Mo, tue, FRIDAY). The time part of start_date is ignored. Example: next_day('2015-01-14', 'TU') = 2015-01-20.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| string                                        | trunc(string<br>date, string<br>format)                                           | Returns date truncated to the unit specified by the format (as of Hive <a href="#">1.2.0</a> ). Supported formats: MONTH/MON/MM, YEAR /YYYY/YY. Example: trunc('2015-03-17', 'MM') = 2015-03-01.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| double                                        | months_between(date1,<br>date2)                                                   | Returns number of months between dates date1 and date2 (as of Hive <a href="#">1.2.0</a> ). If date1 is later than date2, then the result is positive. If date1 is earlier than date2, then the result is negative. If date1 and date2 are either the same days of the month or both last days of months, then the result is always an integer. Otherwise the UDF calculates the fractional portion of the result based on a 31-day month and considers the difference in time components date1 and date2. date1 and date2 type can be date, timestamp or string in the format 'yyyy-MM-dd' or 'yyyy-MM-dd HH:mm:ss'. The result is rounded to 8 decimal places. Example: months_between('1997-02-28 10:30:00', '1996-10-30') = 3.94959677                                                                                                                                                                                                     |

|        |                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------|----------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| string | date_format<br>(date<br>/timestamp<br>/string ts,<br>string pattern) | <p>Converts a date/timestamp/string to a value of string using the specified pattern (as of Hive 1.2.0). The accepted patterns and their behavior depend on the underlying formatter implementation. The pattern argument should be constant.<br/>Example: date_format('2015-04-08', 'y') = '2015'.</p> <p>date_format can be used to implement other UDFs, e.g.:</p> <ul style="list-style-type: none"> <li>dayname(date) is date_format(date, 'EEEE')</li> <li>dayofyear(date) is date_format(date, 'D')</li> </ul> <p>As of Hive 4.0.0 ( <a href="#">HIVE-27673</a> - Getting issue details... <span>STATUS</span> ), the "hive.datetime.formatter" property can be used to control the underlying formatter implementation, and as a consequence the accepted patterns and their behavior. Prior versions always used <a href="https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html">https://docs.oracle.com/javase/8/docs/api/java/text/SimpleDateFormat.html</a> as the underlying formatter.</p> |
|--------|----------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## Conditional Functions

| Return Type | Name(Signature)                                            | Description                                                                                                                     |
|-------------|------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| T           | if(boolean testCondition, T valueTrue, T valueFalseOrNull) | Returns valueTrue when testCondition is true, returns valueFalseOrNull otherwise.                                               |
| boolean     | isnull( a )                                                | Returns true if a is NULL and false otherwise.                                                                                  |
| boolean     | isnotnull ( a )                                            | Returns true if a is not NULL and false otherwise.                                                                              |
| T           | nvl(T value, T default_value)                              | Returns default value if value is null else returns value (as of Hive 0.11).                                                    |
| T           | COALESCE(T v1, T v2, ...)                                  | Returns the first v that is not NULL, or NULL if all v's are NULL.                                                              |
| T           | CASE a WHEN b THEN c [WHEN d THEN e]* [ELSE f] END         | When a = b, returns c; when a = d, returns e; else returns f.                                                                   |
| T           | CASE WHEN a THEN b [WHEN c THEN d]* [ELSE e] END           | When a = true, returns b; when c = true, returns d; else returns e.                                                             |
| T           | nullif( a, b )                                             | Returns NULL if a=b; otherwise returns a (as of Hive 2.3.0).<br><br>Shorthand for: CASE WHEN a = b then NULL else a             |
| void        | assert_true(boolean condition)                             | Throw an exception if 'condition' is not true, otherwise return null (as of Hive 0.8.0). For example, select assert_true (2<1). |

## String Functions

The following built-in String functions are supported in Hive:

| Return Type                  | Name (Signature)                                                          | Description                                                                                                                                                                                                                          |
|------------------------------|---------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| int                          | ascii(string str)                                                         | Returns the numeric value of the first character of str.                                                                                                                                                                             |
| string                       | base64(binary bin)                                                        | Converts the argument from binary to a base 64 string (as of Hive 0.12.0).                                                                                                                                                           |
| int                          | character_length<br>(string str)                                          | Returns the number of UTF-8 characters contained in str (as of Hive 2.2.0). The function char_length is shorthand for this function.                                                                                                 |
| string                       | chr(bigint double A)                                                      | Returns the ASCII character having the binary equivalent to A (as of Hive 1.3.0 and 2.1.0). If A is larger than 256 the result is equivalent to chr(A % 256). Example: select chr(88); returns "X".                                  |
| string                       | concat<br>(string binary A,<br>string binary B...)                        | Returns the string or bytes resulting from concatenating the strings or bytes passed in as parameters in order. For example, concat('foo', 'bar') results in 'foobar'. Note that this function can take any number of input strings. |
| array<struct<string,double>> | context_ngrams<br>(array<array<string>>,<br>array<string>, int K, int pf) | Returns the top-k contextual N-grams from a set of tokenized sentences, given a string of "context". See <a href="#">StatisticsAnd DataMining</a> for more information.                                                              |
| string                       | concat_ws(string SEP, string A, string B...)                              | Like concat() above, but with custom separator SEP.                                                                                                                                                                                  |

|                               |                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------------------------------|---------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| string                        | concat_ws(string SEP, array<string>)                                      | Like concat_ws() above, but taking an array of strings. (as of Hive 0.9.0)                                                                                                                                                                                                                                                                                                                                                                                                 |
| string                        | decode(binary bin, string charset)                                        | Decodes the first argument into a String using the provided character set (one of 'US-ASCII', 'ISO-8859-1', 'UTF-8', 'UTF-16BE', 'UTF-16LE', 'UTF-16'). If either argument is null, the result will also be null. (As of Hive 0.12.0.)                                                                                                                                                                                                                                     |
| string                        | elt(N int, str1 string, str2 string, str3 string, ...)                    | Return string at index number. For example elt(2, 'hello', 'world') returns 'world'. Returns NULL if N is less than 1 or greater than the number of arguments.<br><br>(see <a href="https://dev.mysql.com/doc/refman/5.7/en/string-functions.html#function_elt">https://dev.mysql.com/doc/refman/5.7/en/string-functions.html#function_elt</a> )                                                                                                                           |
| binary                        | encode(string src, string charset)                                        | Encodes the first argument into a BINARY using the provided character set (one of 'US-ASCII', 'ISO-8859-1', 'UTF-8', 'UTF-16BE', 'UTF-16LE', 'UTF-16'). If either argument is null, the result will also be null. (As of Hive 0.12.0.)                                                                                                                                                                                                                                     |
| int                           | field(val T, val1 T, val2 T, val3 T, ...)                                 | Returns the index of val in the val1, val2, val3, ... list or 0 if not found. For example field('world', 'say', 'hello', 'world') returns 3.<br>All primitive types are supported, arguments are compared using str.equals(x). If val is NULL, the return value is 0.<br><br>(see <a href="https://dev.mysql.com/doc/refman/5.7/en/string-functions.html#function_field">https://dev.mysql.com/doc/refman/5.7/en/string-functions.html#function_field</a> )                |
| int                           | find_in_set(string str, string strList)                                   | Returns the first occurrence of str in strList where strList is a comma-delimited string. Returns null if either argument is null. Returns 0 if the first argument contains any commas. For example, find_in_set('ab', 'abc,b,ab,c,def') returns 3.                                                                                                                                                                                                                        |
| string                        | format_number(number x, int d)                                            | Formats the number X to a format like '#,###,###.##', rounded to D decimal places, and returns the result as a string. If D is 0, the result has no decimal point or fractional part. (As of Hive 0.10.0; bug with float types fixed in Hive 0.14.0, decimal type support added in Hive 0.14.0)                                                                                                                                                                            |
| string                        | get_json_object(string json_string, string path)                          | Extracts json object from a json string based on json path specified, and returns json string of the extracted json object. It will return null if the input json string is invalid. <b>NOTE: The json path can only have the characters [0-9a-z_], i.e., no upper-case or special characters. Also, the keys *cannot start with numbers.*</b> This is due to restrictions on Hive column names.                                                                           |
| boolean                       | in_file(string str, string filename)                                      | Returns true if the string str appears as an entire line in filename.                                                                                                                                                                                                                                                                                                                                                                                                      |
| int                           | instr(string str, string substr)                                          | Returns the position of the first occurrence of substr in str. Returns null if either of the arguments are null and returns 0 if substr could not be found in str. Be aware that this is not zero based. The first character in str has index 1.                                                                                                                                                                                                                           |
| int                           | length(string A)                                                          | Returns the length of the string.                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| int                           | locate(string substr, string str[, int pos])                              | Returns the position of the first occurrence of substr in str after position pos.                                                                                                                                                                                                                                                                                                                                                                                          |
| string                        | lower(string A)<br>lcase(string A)                                        | Returns the string resulting from converting all characters of B to lower case. For example, lower('fOoBaR') results in 'foobar'.                                                                                                                                                                                                                                                                                                                                          |
| string                        | lpad(string str, int len, string pad)                                     | Returns str, left-padded with pad to a length of len. If str is longer than len, the return value is shortened to len characters. In case of empty pad string, the return value is null.                                                                                                                                                                                                                                                                                   |
| string                        | ltrim(string A)                                                           | Returns the string resulting from trimming spaces from the beginning(left hand side) of A. For example, ltrim(' foobar ') results in 'foobar'.                                                                                                                                                                                                                                                                                                                             |
| array<struct<string, double>> | ngrams(array<array<string>>, int N, int K, int pf)                        | Returns the top-k N-grams from a set of tokenized sentences, such as those returned by the sentences() UDAF. See <a href="#">StatisticsAndDataMining</a> for more information.                                                                                                                                                                                                                                                                                             |
| int                           | octet_length(string str)                                                  | Returns the number of octets required to hold the string str in UTF-8 encoding (since Hive 2.2.0). Note that octet_length(str) can be larger than character_length(str).                                                                                                                                                                                                                                                                                                   |
| string                        | parse_url(string urlString, string partToExtract [, string keyToExtract]) | Returns the specified part from the URL. Valid values for partToExtract include HOST, PATH, QUERY, REF, PROTOCOL, AUTHORITY, FILE, and USERINFO. For example, parse_url('http://facebook.com/path1/p.php?k1=v1&k2=v2#Ref1', 'HOST') returns 'facebook.com'. Also a value of a particular key in QUERY can be extracted by providing the key as the third argument, for example, parse_url('http://facebook.com/path1/p.php?k1=v1&k2=v2#Ref1', 'QUERY', 'k1') returns 'v1'. |
| string                        | printf(String format, Obj... args)                                        | Returns the input formatted according do printf-style format strings (as of Hive 0.9.0).                                                                                                                                                                                                                                                                                                                                                                                   |

| string               | quote(String text)                                                                           | Returns the quoted string (Includes escape character for any single quotes <a href="#">HIVE-4.0.0</a> ) <table><tr><th>Input</th><th>Output</th></tr><tr><td>NULL</td><td>NULL</td></tr><tr><td>DONT</td><td>'DONT'</td></tr><tr><td>DON'T</td><td>'DON'T'</td></tr></table>                                                                                                                                                                                                                                 | Input | Output | NULL | NULL | DONT | 'DONT' | DON'T | 'DON'T' |
|----------------------|----------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|--------|------|------|------|--------|-------|---------|
| Input                | Output                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |       |        |      |      |      |        |       |         |
| NULL                 | NULL                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |       |        |      |      |      |        |       |         |
| DONT                 | 'DONT'                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |       |        |      |      |      |        |       |         |
| DON'T                | 'DON'T'                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |       |        |      |      |      |        |       |         |
| string               | regexp_extract (string subject, string pattern, int index)                                   | Returns the string extracted using the pattern. For example, regexp_extract('foothebar', 'foo.(.*?)(bar)', 2) returns 'bar.' Note that some care is necessary in using predefined character classes: using 's' as the second argument will match the letter s; '\s' is necessary to match whitespace, etc. The 'index' parameter is the Java regex Matcher group() method index. See <a href="#">docs/api/java/util/regex/Matcher.html</a> for more information on the 'index' or Java regex group() method. |       |        |      |      |      |        |       |         |
| string               | regexp_replace (string INITIAL_STRING, string PATTERN, string REPLACEMENT)                   | Returns the string resulting from replacing all substrings in INITIAL_STRING that match the java regular expression syntax defined in PATTERN with instances of REPLACEMENT. For example, regexp_replace("foobar", "oo ar", "") returns 'fb.' Note that some care is necessary in using predefined character classes: using 's' as the second argument will match the letter s; '\s' is necessary to match whitespace, etc.                                                                                  |       |        |      |      |      |        |       |         |
| string               | repeat(string str, int n)                                                                    | Repeats str n times.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |       |        |      |      |      |        |       |         |
| string               | replace(string A, string OLD, string NEW)                                                    | Returns the string A with all non-overlapping occurrences of OLD replaced with NEW (as of <a href="#">Hive 1.3.0</a> and <a href="#">2.1.0</a> ). Example: select replace("ababab", "abab", "Z"); returns "Zab".                                                                                                                                                                                                                                                                                             |       |        |      |      |      |        |       |         |
| string               | reverse(string A)                                                                            | Returns the reversed string.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |       |        |      |      |      |        |       |         |
| string               | rpadd(string str, int len, string pad)                                                       | Returns str, right-padded with pad to a length of len. If str is longer than len, the return value is shortened to len characters. In case of empty pad string, the return value is null.                                                                                                                                                                                                                                                                                                                    |       |        |      |      |      |        |       |         |
| string               | rtrim(string A)                                                                              | Returns the string resulting from trimming spaces from the end(right hand side) of A. For example, rtrim(' foobar ') results in ' foobar'.                                                                                                                                                                                                                                                                                                                                                                   |       |        |      |      |      |        |       |         |
| array<array<string>> | sentences(string str, string lang, string locale)                                            | Tokenizes a string of natural language text into words and sentences, where each sentence is broken at the appropriate sentence boundary and returned as an array of words. The 'lang' and 'locale' are optional arguments. For example, sentences('Hello there! How are you?') returns ( ("Hello", "there"), ("How", "are", "you") ).                                                                                                                                                                       |       |        |      |      |      |        |       |         |
| string               | space(int n)                                                                                 | Returns a string of n spaces.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |       |        |      |      |      |        |       |         |
| array                | split(string str, string pat)                                                                | Splits str around pat (pat is a regular expression).                                                                                                                                                                                                                                                                                                                                                                                                                                                         |       |        |      |      |      |        |       |         |
| map<string, string>  | str_to_map(text[, delimiter1, delimiter2])                                                   | Splits text into key-value pairs using two delimiters. Delimiter1 separates text into K-V pairs, and Delimiter2 splits each K-V pair. Default delimiters are ',' for delimiter1 and ':' for delimiter2.                                                                                                                                                                                                                                                                                                      |       |        |      |      |      |        |       |         |
| string               | substr (string binary A, int start) substring (string binary A, int start)                   | Returns the substring or slice of the byte array of A starting from start position till the end of string A. For example, substr('foobar', 4) results in 'bar' (see [ <a href="#">http://dev.mysql.com/doc/refman/5.0/en/string-functions.html#function_substr</a> ]).                                                                                                                                                                                                                                       |       |        |      |      |      |        |       |         |
| string               | substr (string binary A, int start, int len) substring (string binary A, int start, int len) | Returns the substring or slice of the byte array of A starting from start position with length len. For example, substr('foobar', 4, 1) results in 'b' (see [ <a href="#">http://dev.mysql.com/doc/refman/5.0/en/string-functions.html#function_substr</a> ]).                                                                                                                                                                                                                                               |       |        |      |      |      |        |       |         |
| string               | substring_index (string A, string delim, int count)                                          | Returns the substring from string A before count occurrences of the delimiter delim (as of <a href="#">Hive 1.3.0</a> ). If count is positive, everything to the left of the final delimiter (counting from the left) is returned. If count is negative, everything to the right of the final delimiter (counting from the right) is returned. Substring_index performs a case-sensitive match when searching for delim. Example: substring_index('www.apache.org', '.', 2) = 'www.apache'.                  |       |        |      |      |      |        |       |         |
| string               | translate (string char varchar input, string char varchar from, string char varchar to)      | Translates the input string by replacing the characters present in the from string with the corresponding characters in the to string. This is similar to the translate function in <a href="#">PostgreSQL</a> . If any of the parameters to this UDF are NULL, the result is NULL as well. (Available as of <a href="#">Hive 0.10.0</a> , for string types)<br><br>Char/varchar support added as of <a href="#">Hive 0.14.0</a> .                                                                           |       |        |      |      |      |        |       |         |
| string               | trim(string A)                                                                               | Returns the string resulting from trimming spaces from both ends of A. For example, trim(' foobar ') results in 'foobar'                                                                                                                                                                                                                                                                                                                                                                                     |       |        |      |      |      |        |       |         |

|        |                                    |                                                                                                                                                                          |
|--------|------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| binary | unbase64(string str)               | Converts the argument from a base 64 string to BINARY. (As of Hive <a href="#">0.12.0</a> .)                                                                             |
| string | upper(string A)<br>ucase(string A) | Returns the string resulting from converting all characters of A to upper case. For example, upper('fOoBaR') results in 'FOOBAR'.                                        |
| string | initcap(string A)                  | Returns string, with the first letter of each word in uppercase, all other letters in lowercase. Words are delimited by whitespace. (As of Hive <a href="#">1.1.0</a> .) |
| int    | levenshtein(string A, string B)    | Returns the Levenshtein distance between two strings (as of Hive <a href="#">1.2.0</a> ). For example, levenshtein('kitten', 'sitting') results in 3.                    |
| string | soundex(string A)                  | Returns soundex code of the string (as of Hive <a href="#">1.2.0</a> ). For example, soundex('Miller') results in M460.                                                  |

## Data Masking Functions

The following built-in data masking functions are supported in Hive:

| Return Type | Name (Signature)                                                 | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------|------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| string      | mask(string str[, string upper[, string lower[, string number]]) | Returns a masked version of str (as of Hive <a href="#">2.1.0</a> ). By default, upper case letters are converted to "X", lower case letters are converted to "x" and numbers are converted to "n". For example mask("abcd-EFGH-8765-4321") results in xxxx-XXXX-nnnn-nnnn. You can override the characters used in the mask by supplying additional arguments: the second argument controls the mask character for upper case letters, the third argument for lower case letters and the fourth argument for numbers. For example, mask("abcd-EFGH-8765-4321", "U", "l", "#") results in llll-UUUU-####-####. |
| string      | mask_first_n(string str[, int n])                                | Returns a masked version of str with the first n values masked (as of Hive <a href="#">2.1.0</a> ). Upper case letters are converted to "X", lower case letters are converted to "x" and numbers are converted to "n". For example, mask_first_n("1234-5678-8765-4321", 4) results in nnnn-5678-8765-4321.                                                                                                                                                                                                                                                                                                     |
| string      | mask_last_n(string str[, int n])                                 | Returns a masked version of str with the last n values masked (as of Hive <a href="#">2.1.0</a> ). Upper case letters are converted to "X", lower case letters are converted to "x" and numbers are converted to "n". For example, mask_last_n("1234-5678-8765-4321", 4) results in 1234-5678-8765-nnnn.                                                                                                                                                                                                                                                                                                       |
| string      | mask_show_first_n(string str[, int n])                           | Returns a masked version of str, showing the first n characters unmasked (as of Hive <a href="#">2.1.0</a> ). Upper case letters are converted to "X", lower case letters are converted to "x" and numbers are converted to "n". For example, mask_show_first_n("1234-5678-8765-4321", 4) results in 1234-nnnn-nnnn-nnnn.                                                                                                                                                                                                                                                                                      |
| string      | mask_show_last_n(string str[, int n])                            | Returns a masked version of str, showing the last n characters unmasked (as of Hive <a href="#">2.1.0</a> ). Upper case letters are converted to "X", lower case letters are converted to "x" and numbers are converted to "n". For example, mask_show_last_n("1234-5678-8765-4321", 4) results in nnnn-nnnn-nnnn-4321.                                                                                                                                                                                                                                                                                        |
| string      | mask_hash(string char varchar str)                               | Returns a hashed value based on str (as of Hive <a href="#">2.1.0</a> ). The hash is consistent and can be used to join masked values together across tables. This function returns null for non-string types.                                                                                                                                                                                                                                                                                                                                                                                                 |

## Misc. Functions

| Return Type | Name (Signature)                              | Description                                                                                                                                                                                                                                                            |
|-------------|-----------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| varies      | java_method(class, method[, arg1[, arg2...]]) | Synonym for reflect. (As of Hive <a href="#">0.9.0</a> .)                                                                                                                                                                                                              |
| varies      | reflect(class, method[, arg1[, arg2...]])     | Calls a Java method by matching the argument signature, using reflection. (As of Hive <a href="#">0.7.0</a> .) See <a href="#">Reflect (Generic) UDF</a> for examples.                                                                                                 |
| int         | hash(a1[, a2...])                             | Returns a hash value of the arguments. (As of Hive 0.4.)                                                                                                                                                                                                               |
| string      | current_user()                                | Returns current user name from the configured authenticator manager (as of Hive <a href="#">1.2.0</a> ). Could be the same as the user provided when connecting, but with some authentication managers (for example HadoopDefaultAuthenticator) it could be different. |
| string      | logged_in_user()                              | Returns current user name from the session state (as of Hive <a href="#">2.2.0</a> ). This is the username provided when connecting to Hive.                                                                                                                           |
| string      | current_database()                            | Returns current database name (as of Hive <a href="#">0.13.0</a> ).                                                                                                                                                                                                    |

|        |                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|--------|----------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| string | md5(string<br>/binary)                                         | Calculates an MD5 128-bit checksum for the string or binary (as of Hive <a href="#">1.3.0</a> ). The value is returned as a string of 32 hex digits, or NULL if the argument was NULL. Example: md5('ABC') = '902fbdd2b1df0c4f70b4a5d23525e932'.                                                                                                                                                                                                                                                                                                                                                  |
| string | sha1(string<br>/binary)<br><br>sha(string<br>/binary)          | Calculates the SHA-1 digest for string or binary and returns the value as a hex string (as of Hive <a href="#">1.3.0</a> ). Example: sha1('ABC') = '3c01bdbb26f358bab27f267924aa2c9a03fcfdb8'.                                                                                                                                                                                                                                                                                                                                                                                                    |
| bigint | crc32(string<br>/binary)                                       | Computes a cyclic redundancy check value for string or binary argument and returns bigint value (as of Hive <a href="#">1.3.0</a> ). Example: crc32('ABC') = 2743272264.                                                                                                                                                                                                                                                                                                                                                                                                                          |
| string | sha2(string<br>/binary, int)                                   | Calculates the SHA-2 family of hash functions (SHA-224, SHA-256, SHA-384, and SHA-512) (as of Hive <a href="#">1.3.0</a> ). The first argument is the string or binary to be hashed. The second argument indicates the desired bit length of the result, which must have a value of 224, 256, 384, 512, or 0 (which is equivalent to 256). SHA-224 is supported starting from Java 8. If either argument is NULL or the hash length is not one of the permitted values, the return value is NULL. Example: sha2('ABC', 256) = 'b5d4045c3f466fa91fe2cc6abe79232a1a57cdf104f7a26e716e0a1e2789df78'. |
| binary | aes_encrypt<br>(input string<br>/binary, key<br>string/binary) | Encrypt input using AES (as of Hive <a href="#">1.3.0</a> ). Key lengths of 128, 192 or 256 bits can be used. 192 and 256 bits keys can be used if Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction Policy Files are installed. If either argument is NULL or the key length is not one of the permitted values, the return value is NULL. Example: base64(aes_encrypt('ABC', '1234567890123456')) = 'y6Ss+zCYObpCbgfWfyNWTw=='.                                                                                                                                                 |
| binary | aes_decrypt<br>(input binary,<br>key string<br>/binary)        | Decrypt input using AES (as of Hive <a href="#">1.3.0</a> ). Key lengths of 128, 192 or 256 bits can be used. 192 and 256 bits keys can be used if Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction Policy Files are installed. If either argument is NULL or the key length is not one of the permitted values, the return value is NULL. Example: aes_decrypt(unbase64('y6Ss+zCYObpCbgfWfyNWTw=='), '1234567890123456') = 'ABC'.                                                                                                                                               |
| string | version()                                                      | Returns the Hive version (as of Hive <a href="#">2.1.0</a> ). The string contains 2 fields, the first being a build number and the second being a build hash. Example: "select version();" might return "2.1.0.2.5.0.0-1245r027527b9c5ce1a3d7d0b6d2e6de2378fb0c39232". Actual results will depend on your build.                                                                                                                                                                                                                                                                                  |
| bigint | surrogate_key<br>([write_id_bits<br>,<br>task_id_bits])        | Automatically generate numerical Ids for rows as you enter data into a table. Can only be used as default value for acid or insert-only tables.                                                                                                                                                                                                                                                                                                                                                                                                                                                   |

## xpath

The following functions are described in [LanguageManual XPathUDF](#):

- xpath, xpath\_short, xpath\_int, xpath\_long, xpath\_float, xpath\_double, xpath\_number, xpath\_string

## get\_json\_object

A limited version of JSONPath is supported:

- \$ : Root object
- .. : Child operator
- [] : Subscript operator for array
- \* : Wildcard for []

Syntax not supported that's worth noticing:

- : Zero length string as key
- .. : Recursive descent
- @ : Current object/element
- () : Script expression
- ?() : Filter (script) expression.
- [] : Union operator
- [start:end.step] : array slice operator

Example: src\_json table is a single column (json), single row table:

```
+----+
                        json
+----+
{"store":
  {"fruit":\[{{"weight":8,"type":"apple"},{"weight":9,"type":"pear"}],
   "bicycle":{"price":19.95,"color":"red"}
 },
 "email":"amy@only_for_json_udf_test.net",
 "owner":"amy"
}
+----+
```

The fields of the json object can be extracted using these queries:

```
hive> SELECT get_json_object(src_json.json, '$.owner') FROM src_json;
amy

hive> SELECT get_json_object(src_json.json, '$.store.fruit\[0]') FROM src_json;
{"weight":8,"type":"apple"}

hive> SELECT get_json_object(src_json.json, '$.non_exist_key') FROM src_json;
NULL
```

## Built-in Aggregate Functions (UDAF)

The following built-in aggregate functions are supported in Hive:

| Return Type | Name (Signature)                                        | Description                                                                                                                                                                                                                                                                                                                                                                                                          |
|-------------|---------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| BIGINT      | count(*), count(expr), count(DISTINCT expr [, expr...]) | count(*) - Returns the total number of retrieved rows, including rows containing NULL values.<br><br>count(expr) - Returns the number of rows for which the supplied expression is non-NULL.<br><br>count(DISTINCT expr[, expr]) - Returns the number of rows for which the supplied expression(s) are unique and non-NULL. Execution of this can be optimized with <a href="#">hive.optimize.distinct.rewrite</a> . |
| DOUBLE      | sum(col), sum(DISTINCT col)                             | Returns the sum of the elements in the group or the sum of the distinct values of the column in the group.                                                                                                                                                                                                                                                                                                           |
| DOUBLE      | avg(col), avg(DISTINCT col)                             | Returns the average of the elements in the group or the average of the distinct values of the column in the group.                                                                                                                                                                                                                                                                                                   |
| DOUBLE      | min(col)                                                | Returns the minimum of the column in the group.                                                                                                                                                                                                                                                                                                                                                                      |
| DOUBLE      | max(col)                                                | Returns the maximum value of the column in the group.                                                                                                                                                                                                                                                                                                                                                                |
| DOUBLE      | variance(col), var_pop(col)                             | Returns the variance of a numeric column in the group.                                                                                                                                                                                                                                                                                                                                                               |
| DOUBLE      | var_samp(col)                                           | Returns the unbiased sample variance of a numeric column in the group.                                                                                                                                                                                                                                                                                                                                               |
| DOUBLE      | stddev_pop(col)                                         | Returns the standard deviation of a numeric column in the group.                                                                                                                                                                                                                                                                                                                                                     |
| DOUBLE      | stddev_samp(col)                                        | Returns the unbiased sample standard deviation of a numeric column in the group.                                                                                                                                                                                                                                                                                                                                     |
| DOUBLE      | covar_pop(col1, col2)                                   | Returns the population covariance of a pair of numeric columns in the group.                                                                                                                                                                                                                                                                                                                                         |
| DOUBLE      | covar_samp(col1, col2)                                  | Returns the sample covariance of a pair of a numeric columns in the group.                                                                                                                                                                                                                                                                                                                                           |
| DOUBLE      | corr(col1, col2)                                        | Returns the Pearson coefficient of correlation of a pair of a numeric columns in the group.                                                                                                                                                                                                                                                                                                                          |
| DOUBLE      | percentile (BIGINT col, p)                              | Returns the exact p <sup>th</sup> percentile of a column in the group (does not work with floating point types). p must be between 0 and 1. NOTE: A true percentile can only be computed for integer values. Use PERCENTILE_APPROX if your input is non-integral.                                                                                                                                                    |

|                          |                                                                                  |                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| array<double>            | percentile (BIGINT col, array(p <sub>1</sub> [, p <sub>2</sub> ]...))            | Returns the exact percentiles p <sub>1</sub> , p <sub>2</sub> , ... of a column in the group (does not work with floating point types). p <sub>i</sub> must be between 0 and 1. NOTE: A true percentile can only be computed for integer values. Use PERCENTILE_APPROX if your input is non-integral.                                                                |
| DOUBLE                   | percentile_approx(DOUBLE col, p [, B])                                           | Returns an approximate p <sup>th</sup> percentile of a numeric column (including floating point types) in the group. The B parameter controls approximation accuracy at the cost of memory. Higher values yield better approximations, and the default is 10,000. When the number of distinct values in col is smaller than B, this gives an exact percentile value. |
| array<double>            | percentile_approx(DOUBLE col, array(p <sub>1</sub> [, p <sub>2</sub> ]... [, B]) | Same as above, but accepts and returns an array of percentile values instead of a single one.                                                                                                                                                                                                                                                                        |
| double                   | regr_avgx (independent, dependent)                                               | Equivalent to avg(dependent). As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                                                                                                                                     |
| double                   | regr_avgy (independent, dependent)                                               | Equivalent to avg(independent). As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                                                                                                                                   |
| double                   | regr_count (independent, dependent)                                              | Returns the number of non-null pairs used to fit the linear regression line. As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                                                                                      |
| double                   | regr_intercept (independent, dependent)                                          | Returns the y-intercept of the <a href="#">linear regression line</a> , i.e. the value of b in the equation dependent = a * independent + b. As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                      |
| double                   | regr_r2 (independent, dependent)                                                 | Returns the <a href="#">coefficient of determination</a> for the regression. As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                                                                                      |
| double                   | regr_slope (independent, dependent)                                              | Returns the slope of the <a href="#">linear regression line</a> , i.e. the value of a in the equation dependent = a * independent + b. As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                            |
| double                   | regr_sxx (independent, dependent)                                                | Equivalent to regr_count(independent, dependent) * var_pop(dependent). As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                                                                                            |
| double                   | regr_sxy (independent, dependent)                                                | Equivalent to regr_count(independent, dependent) * covar_pop(independent, dependent). As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                                                                             |
| double                   | regr_syy (independent, dependent)                                                | Equivalent to regr_count(independent, dependent) * var_pop(independent). As of <a href="#">Hive 2.2.0</a> .                                                                                                                                                                                                                                                          |
| array<struct {'x', 'y'}> | histogram_numeric(col, b)                                                        | Computes a histogram of a numeric column in the group using b non-uniformly spaced bins. The output is an array of size b of double-valued (x,y) coordinates that represent the bin centers and heights                                                                                                                                                              |
| array                    | collect_set(col)                                                                 | Returns a set of objects with duplicate elements eliminated.                                                                                                                                                                                                                                                                                                         |
| array                    | collect_list(col)                                                                | Returns a list of objects with duplicates. (As of Hive <a href="#">0.13.0</a> .)                                                                                                                                                                                                                                                                                     |
| INTEGER                  | ntile(INTEGER x)                                                                 | Divides an ordered partition into x groups called buckets and assigns a bucket number to each row in the partition. This allows easy calculation of tertiles, quartiles, deciles, percentiles and other common summary statistics. (As of Hive <a href="#">0.11.0</a> .)                                                                                             |

## Built-in Table-Generating Functions (UDTF)

Normal user-defined functions, such as concat(), take in a single input row and output a single output row. In contrast, table-generating functions transform a single input row to multiple output rows.

| Row-set columns types                 | Name (Signature)                                       | Description                                                                                                                                                                  |
|---------------------------------------|--------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| T                                     | explode (ARRAY<T> a)                                   | Explodes an array to multiple rows. Returns a row-set with a single column (col), one row for each element from the array.                                                   |
| T <sub>key</sub> , T <sub>value</sub> | explode(MAP<T <sub>key</sub> , T <sub>value</sub> > m) | Explodes a map to multiple rows. Returns a row-set with a two columns (key,value) , one row for each key-value pair from the input map. (As of Hive <a href="#">0.8.0</a> .) |

|                                              |                                                                                                  |                                                                                                                                                                                                                                                                                       |
|----------------------------------------------|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| int,T                                        | posexplode(ARRAY<T> a)                                                                           | Explodes an array to multiple rows with additional positional column of <i>int</i> type (position of items in the original array, starting with 0). Returns a row-set with two columns ( <i>pos,val</i> ), one row for each element from the array.                                   |
| T <sub>1</sub> ,...,T <sub>n</sub>           | inline<br>(ARRAY<STRUCT<f <sub>1</sub> :T <sub>1</sub> ,...,f <sub>n</sub> :T <sub>n</sub> >> a) | Explodes an array of structs to multiple rows. Returns a row-set with N columns (N = number of top level elements in the struct), one row per struct from the array. (As of Hive <a href="#">0.10</a> .)                                                                              |
| T <sub>1</sub> ,...,T <sub>n/r</sub>         | stack(int r,T <sub>1</sub> V <sub>1</sub> ...,T <sub>n/r</sub> V <sub>n</sub> )                  | Breaks up <i>n</i> values V <sub>1</sub> ,...,V <sub>n</sub> into <i>r</i> rows. Each row will have <i>n/r</i> columns. <i>r</i> must be constant.                                                                                                                                    |
|                                              |                                                                                                  |                                                                                                                                                                                                                                                                                       |
| string <sub>1</sub> ,...,string <sub>n</sub> | json_tuple(string jsonStr,string k <sub>1</sub> ,...,string k <sub>n</sub> )                     | Takes JSON string and a set of <i>n</i> keys, and returns a tuple of <i>n</i> values. This is a more efficient version of the get_json_object UDF because it can get multiple keys with just one call.                                                                                |
| string <sub>1</sub> ,...,string <sub>n</sub> | parse_url_tuple(string urlStr,string p <sub>1</sub> ,...,string p <sub>n</sub> )                 | Takes URL string and a set of <i>n</i> URL parts, and returns a tuple of <i>n</i> values. This is similar to the parse_url() UDF but can extract multiple parts at once out of a URL. Valid part names are: HOST, PATH, QUERY, REF, PROTOCOL, AUTHORITY, FILE, USERINFO, QUERY:<KEY>. |

## Usage Examples

### explode (array)

```
select explode(array('A','B','C'));
select explode(array('A','B','C')) as col;
select tf.* from (select 0) t lateral view explode(array('A','B','C')) tf;
select tf.* from (select 0) t lateral view explode(array('A','B','C')) tf as col;
```

### explode (map)

```
select explode(map('A',10,'B',20,'C',30));
select explode(map('A',10,'B',20,'C',30)) as (key,value);
select tf.* from (select 0) t lateral view explode(map('A',10,'B',20,'C',30)) tf;
select tf.* from (select 0) t lateral view explode(map('A',10,'B',20,'C',30)) tf as key,value;
```

### posexplode (array)

```
select posexplode(array('A','B','C'));
select posexplode(array('A','B','C')) as (pos,val);
select tf.* from (select 0) t lateral view posexplode(array('A','B','C')) tf;
select tf.* from (select 0) t lateral view posexplode(array('A','B','C')) tf as pos,val;
```

### inline (array of structs)

```
select inline(array(struct('A',10,date '2015-01-01'),struct('B',20,date '2016-02-02')));
select inline(array(struct('A',10,date '2015-01-01'),struct('B',20,date '2016-02-02'))) as (col1,col2,col3);
select tf.* from (select 0) t lateral view inline(array(struct('A',10,date '2015-01-01'),struct('B',20,date '2016-02-02'))) tf;
select tf.* from (select 0) t lateral view inline(array(struct('A',10,date '2015-01-01'),struct('B',20,date '2016-02-02'))) tf as col1,col2,col3;
```

### stack (values)

```
select stack(2,'A',10,date '2015-01-01','B',20,date '2016-01-01');
select stack(2,'A',10,date '2015-01-01','B',20,date '2016-01-01') as (col0,col1,col2);
select tf.* from (select 0) t lateral view stack(2,'A',10,date '2015-01-01','B',20,date '2016-01-01') tf;
select tf.* from (select 0) t lateral view stack(2,'A',10,date '2015-01-01','B',20,date '2016-01-01') tf as col0,col1,col2;
```

Using the syntax "SELECT udtf(col) AS colAlias..." has a few limitations:

- No other expressions are allowed in SELECT
  - SELECT pageid, explode(adid\_list) AS myCol... is not supported
- UDTF's can't be nested
  - SELECT explode(explode(adid\_list)) AS myCol... is not supported
- GROUP BY / CLUSTER BY / DISTRIBUTE BY / SORT BY is not supported
  - SELECT explode(adid\_list) AS myCol ... GROUP BY myCol is not supported

Please see [LanguageManual LateralView](#) for an alternative syntax that does not have these limitations.

Also see [Writing UDTFs](#) if you want to create a custom UDTF.

### explode

`explode()` takes in an array (or a map) as an input and outputs the elements of the array (map) as separate rows. UDTFs can be used in the SELECT expression list and as a part of LATERAL VIEW.

As an example of using `explode()` in the SELECT expression list, consider a table named `myTable` that has a single column (`myCol`) and two rows:

| Array<int> myCol |
|------------------|
| [100,200,300]    |
| [400,500,600]    |

Then running the query:

```
SELECT explode(myCol) AS myNewCol FROM myTable;
```

will produce:

| (int) myNewCol |
|----------------|
| 100            |
| 200            |
| 300            |
| 400            |
| 500            |
| 600            |

The usage with Maps is similar:

```
SELECT explode(myMap) AS (myMapKey, myMapValue) FROM myMapTable;
```

## posexplode



### Version

Available as of Hive 0.13.0. See [HIVE-4943](#).

`posexplode()` is similar to `explode` but instead of just returning the elements of the array it returns the element as well as its position in the original array.

As an example of using `posexplode()` in the `SELECT` expression list, consider a table named `myTable` that has a single column (`myCol`) and two rows:

| Array<int> myCol |
|------------------|
| [100,200,300]    |
| [400,500,600]    |

Then running the query:

```
SELECT posexplode(myCol) AS pos, myNewCol FROM myTable;
```

will produce:

| (int) pos | (int) myNewCol |
|-----------|----------------|
| 1         | 100            |
| 2         | 200            |
| 3         | 300            |
| 1         | 400            |
| 2         | 500            |
| 3         | 600            |

## json\_tuple

A new `json_tuple()` UDTF is introduced in Hive 0.7. It takes a set of names (keys) and a JSON string, and returns a tuple of values using one function. This is much more efficient than calling `GET_JSON_OBJECT` to retrieve more than one key from a single JSON string. In any case where a single JSON string would be parsed more than once, your query will be more efficient if you parse it once, which is what `JSON_TUPLE` is for. As `JSON_TUPLE` is a UDTF, you will need to use the [LATERAL VIEW](#) syntax in order to achieve the same goal.

For example,

```
select a.timestamp, get_json_object(a.appevents, '$.eventid'), get_json_object(a.appenvets, '$.eventname') from
log a;
```

should be changed to:

```
select a.timestamp, b.*
from log a lateral view json_tuple(a.appevent, 'eventid', 'eventname') b as f1, f2;
```

## parse\_url\_tuple

The `parse_url_tuple()` UDTF is similar to `parse_url()`, but can extract multiple parts of a given URL, returning the data in a tuple. Values for a particular key in `QUERY` can be extracted by appending a colon and the key to the `partToExtract` argument, for example, `parse_url_tuple('http://facebook.com/path1/p.php?k1=v1&k2=v2#Ref1', 'QUERY:k1', 'QUERY:k2')` returns a tuple with values of 'v1','v2'. This is more efficient than calling `parse_url()` multiple times. All the input parameters and output column types are string.

```
SELECT b.*
FROM src LATERAL VIEW parse_url_tuple(fullurl, 'HOST', 'PATH', 'QUERY', 'QUERY:id') b as host, path, query,
query_id LIMIT 1;
```

## GROUPing and SORTing on f(column)

A typical OLAP pattern is that you have a timestamp column and you want to group by daily or other less granular date windows than by second. So you might want to select `concat(year(dt),month(dt))` and then group on that `concat()`. But if you attempt to `GROUP BY` or `SORT BY` a column on which you've applied a function and alias, like this:

```
select f(col) as fc, count(*) from table_name group by fc;
```

you will get an error:

```
FAILED: Error in semantic analysis: line 1:69 Invalid Table Alias or Column Reference fc
```

because you are not able to `GROUP BY` or `SORT BY` a column alias on which a function has been applied. There are two workarounds. First, you can reformulate this query with subqueries, which is somewhat complicated:

```
select sq.fc,col1,col2,...,colN,count(*) from
(select f(col) as fc,col1,col2,...,colN from table_name) sq
group by sq.fc,col1,col2,...,colN;
```

Or you can make sure not to use a column alias, which is simpler:

```
select f(col) as fc, count(*) from table_name group by f(col);
```

Contact Tim Ellis (tellis) at RiotGames dot com if you would like to discuss this in further detail.

## Utility Functions

| Function Name | Return Type | Description                                                   | To Run                              |
|---------------|-------------|---------------------------------------------------------------|-------------------------------------|
| version       | String      | Provides the Hive version Details (Package built version)     | <code>select version();</code>      |
| buildversion  | String      | Extension of the Version function which includes the checksum | <code>select buildversion();</code> |

## UDF internals

The context of a UDF's evaluate method is one row at a time. A simple invocation of a UDF like

```
SELECT length(string_col) FROM table_name;
```

would evaluate the length of each of the string\_col's values in the map portion of the job. The side effect of the UDF being evaluated on the map-side is that you can't control the order of rows which get sent to the mapper. It is the same order in which the file split sent to the mapper gets deserialized. Any reduce side operation (such as SORT BY, ORDER BY, regular JOIN, etc.) would apply to the UDFs output as if it is just another column of the table. This is fine since the context of the UDF's evaluate method is meant to be one row at a time.

If you would like to control which rows get sent to the same UDF (and possibly in what order), you will have the urge to make the UDF evaluate during the reduce phase. This is achievable by making use of [DISTRIBUTE BY](#), [DISTRIBUTE BY + SORT BY](#), [CLUSTER BY](#). An example query would be:

```
SELECT reducer_udf(my_col, distribute_col, sort_col) FROM
(SELECT my_col, distribute_col, sort_col FROM table_name DISTRIBUTE BY distribute_col SORT BY distribute_col,
sort_col) t
```

However, one could argue that the very premise of your requirement to control the set of rows sent to the same UDF is to do aggregation in that UDF. In such a case, using a User Defined Aggregate Function (UDAF) is a better choice. You can read more about writing a UDAF [here](#). Alternatively, you can use a custom reduce script to accomplish the same using [Hive's Transform functionality](#). Both of these options would do aggregations on the reduce side.

## Creating Custom UDFs

For information about how to create a custom UDF, see [Hive Plugins](#) and [Create Function](#).

```
select explode(array('A','B','C'));select explode(array('A','B','C')) as col;select tf.* from (select 0) t lateral view explode(array('A','B','C')) tf;select tf.* from
(select 0) t lateral view explode(array('A','B','C')) tf as col;
```