Databend

Performance tuning in databend

sundyli@datafuselabs（李本旺）
http://github.com/sundy-li/

目录
CONTENT

1 Knowledges before tuning the codes

2 Performance tuning in practice

3 Pipeline/Dataflow executor

4 WIP & TODO

1 Knowledges before tuning the codes

memory hierarchy

From CSAPP

Latency Numbers

[Numbers Every Programmer Should Know By Year (colin-scott.github.io)]

Principle of locality

1. Data Locality
2. Time Locality

Data locality

Row-major or Column-major

```
int sum_array_rows(int a[M][N])
{
    int i, j, sum = 0;

    for (i = 0; i < M; i++)
        for (j = 0; j < N; j++)
            sum += a[i][j];
    return sum;
}
```

```
int sum_array_cols(int a[M][N])
{
    int i, j, sum = 0;

    for (j = 0; j < N; j++)
        for (i = 0; i < M; i++)
            sum += a[i][j];
    return sum;
}
```

Expressing Performance


```
/* Compute prefix sum of vector a */
void psum1(float a[], float p[], long n)
{
    long i;
    p[0] = a[0];
    for (i = 1; i < n; i++)
        p[i] = p[i-1] + a[i];
}

void psum2(float a[], float p[], long n)
{
    long i;
    p[0] = a[0];
    for (i = 1; i < n-1; i+=2) {
        float mid_val = p[i-1] + a[i];
        p[i]     = mid_val;
        p[i+1]   = mid_val + a[i+1];
    }
    /* For even n, finish remaining element */
    if (i < n)
        p[i] = p[i-1] + a[i];
}
```

Figure 5.1 **Prefix-sum functions.** These functions provide examples for how we express program performance.

CPE

From CSAPP

Tools

Some useful tools & scripts in tuning databend's performance

1. Print & Instant elapsed
2. Perf is a good tool, pprof, flamegraph
3. cargo bench & criteron
4. databend-benchmark
5. numbers table && ignore function && progress

```
select count() from numbers_mt(10000000) where not ignore(toString(number))
```

2 Performance tuning in practice

inline vs never inline

Rust hint: #[inline], #[inline(never)], #[inline(always)]

1. Inlining is an optimizing transformation which replaces a call to a function with its body.
2. Private functions may not need inline due to LLVM's heuristic inline.
3. Generic functions are implicitly inlinable.

Yet:

1. Make compile times worser with larger target size.
2. Generic functions may make cause code bloat with inline

```
// Public, generic function.
// Will cause code bloat if not handled carefully!
pub fn parse_str(wat: impl AsRef<str>) -> Result<Vec<u8>> {
  // Immediately delegate to a non-generic function.
  _parse_str(wat.as_ref())
}

// Separate-compilation friendly private implementation.
fn _parse_str(wat: &str) -> Result<Vec<u8>> {
    ...
}
```

inline vs never inline

Case 1: inline vs never inline

```rust
fn default_crc32(array: &PrimitiveArray<u32>) -> u32 {
    array.values().iter().fold(init: 0, f: |acc: u32, x: &u32| {
        let mut hasher: Hasher = crc32fast::Hasher::new();
        hasher.update(buf: &[*x as u8]);
        let checksum: u32 = hasher.finalize();
        acc + checksum
    })
}

fn inlined_crc32(array: &PrimitiveArray<u32>) -> u32 {
    array.values().iter().fold(init: 0, f: |acc: u32, x: &u32| {
        let mut hasher: Hasher = mycrc32::Hasher::new();
        hasher.update(buf: &[*x as u8]);
        let checksum: u32 = hasher.finalize();
        acc + checksum
    })
}
```

default_crc32 time: [16.184 ms 16.196 ms 16.209 ms]

inlined_crc32 time: [1.6317 ms 1.6353 ms 1.6396 ms]

10X improvement

inline is your friend!!

inline vs never inline

Case 2: inline vs never inline

```rust
impl<Method: HashMethod + PolymorphicKeysHelper<Method>> Aggregator<Method> {
    pub fn create(method: Method, params: AggregatorParamsRef) -> Aggregator<Method> {
        Aggregator { method, params }
    }

    // If we set it to inline(performance degradation).
    // Because it will make other internal functions to no inline
    #[inline(never)]        zhang2014, 3 months ago • Try to determine inline
    pub async fn aggregate(…
    ) -> Result<Method::State> {
        // This may be confusing
        // It will help us improve performance ~10% when we declare local references for them.
        let hash_method: &Method = &self.method;
        let aggregator_params: &AggregatorParams = self.params.as_ref();

        let mut state: <Method as PolymorphicKeysHelper<…>>::State = hash_method.aggregate_state();

        match aggregator_params.aggregate_functions.is_empty() {…

        Ok(state)
    }
```

never inline is your friend!!!

codes in databend

No extra allocation/copy

Case 3: numbers stream in arrow format

select sum(number) from numbers_mt(10000000000)

```
-            let data: Vec<u64> = (current.begin..current.end).collect();
-            let block =
-                DataBlock::create(self.schema.clone(), vec![Arc::new(UInt64Array::from(data))]);
-            Some(Ok(block))
```

```
+            unsafe {
+                let layout = Layout::from_size_align_unchecked(
+                    length * mem::size_of::<u64>(),
+                    mem::size_of::<u64>(),
+                );
+                let p = std::alloc::alloc(layout) as *mut u64;
+                for i in current.begin..current.end {
+                    *p.offset((i - current.begin) as isize) = i;
+                }
+                let buffer =
+                    Buffer::from_raw_parts(NonNull::new(p as *mut u8).unwrap(), length, length);
+
+                let arr_data = ArrayData::builder(DataType::UInt64)
+                    .len(length)
+                    .offset(0)
+                    .add_buffer(buffer)
+                    .build();
+
+                let block = DataBlock::create(
+                    self.schema.clone(),
+                    vec![Arc::new(UInt64Array::from(arr_data))],
+                );
+                Some(Ok(block))
+            }
```

Databend PR #19

No extra allocation/copy

Case 3: numbers stream in arrow format

Profile and tunning

- Count, sum(top3) = 86.44%

```
31.70%  tokio-runtime-w  libc-2.19.so        [.] __memcpy_sse2_unaligned
27.95%  tokio-runtime-w  libc-2.19.so        [.] memset
26.79%  tokio-runtime-w  fuse-query          [.] <alloc::vec::Vec<T> as alloc::vec::spec_from_iter::SpecFr
 1.69%  tokio-runtime-w  libc-2.19.so        [.] malloc
```

- Max, sum(top4) = 87.31%

```
28.34%  tokio-runtime-w  libc-2.19.so       [.] __memcpy_sse2_unaligned
24.52%  tokio-runtime-w  fuse-query         [.] <alloc::vec::Vec<T> as alloc::vec::spec_from_iter::SpecFrc
24.20%  tokio-runtime-w  libc-2.19.so       [.] memset
10.25%  tokio-runtime-w  fuse-query         [.] arrow::compute::kernels::aggregate::max
 1.60
```

As we can see, `max` query has about `10.25%` extra cpu costs in `arrow::compute::kernels::aggregate::max`.
But why it performances the same as the `count` query?

After some thinking, I realized that must be the problem of source data generation, the speed of upstream generation is slow, so the downstream are hungry. Just like two men are waiting for food, one of them can finish the dinner in 5 minutes, the other may take 1 minute, but food preparation time may take an hour, so we see they just finish the dinner at the same time.

memcpy is the hot path via perf

No extra allocation/copy

Case 4.1: cast primitive to string

select count() from numbers_mt(10000000000) where not ignore(toString(number))

```diff
@@ -19,9 +17,21 @@
 pub fn primitive_to_binary<T: NativeType + lexical_core::ToLexical, O: Offset>(
     from: &PrimitiveArray<T>,
 ) -> BinaryArray<O> {
-    let iter = from.iter().map(|x| x.map(|x| lexical_to_bytes(*x)));
-
-    BinaryArray::from_trusted_len_iter(iter)
+    let mut buffer = vec![];
+    let builder = from.iter().fold(
+        MutableBinaryArray::<O>::with_capacity(from.len()),
+        |mut builder, x| {
+            match x {
+                Some(x) => {
+                    lexical_to_bytes_mut(*x, &mut buffer);
+                    builder.push(Some(buffer.as_slice()));
+                }
+                None => builder.push_null(),
+            }
+            builder
+        },
+    );
+    builder.into()
 }
```

[Arrow2 PR #646]

Improved performance in cast Primitive to Binary/String (2x) #646

Merged jorgecarleitao merged 2 commits into jorgecarleitao:main from datafuse-extras:primitive-cast 2 days ago

Conversation 2 | Commits 2 | Checks 24 | Files changed 2

sundy-li commented 2 days ago • edited Contributor

Reuse the temp bytes buffer.

```
cast int32 to binary 512
                time:   [10.650 us 10.761 us 10.885 us]
                change: [-47.175% -46.939% -46.649%] (p = 0.00 < 0.05)
                Performance has improved.
```

3

No extra allocation/copy

Case 4.2: cast primitive to string no extra

select count() from numbers_mt(10000000000) where not ignore(toString(number))

```rust
    /// Write values directly into the underlying data.
    /// # Safety
    /// Caller must ensure that `len <= self.capacity()`
    #[inline]
    pub unsafe fn write_values<F>(&mut self, mut f: F)
    where
        F: FnMut(&mut [u8]) -> usize,
    {
        // ensure values has enough capacity and size to write
        self.values.set_len(self.values.capacity());
        let buffer = &mut self.values.as_mut_slice()
[self.offsets.last().unwrap().to_usize()..];
        let len = f(buffer);
        let new_len = self.offsets.last().unwrap().to_usize() + len;
        self.values.set_len(new_len);

        let size = O::from_usize(new_len).ok_or(ArrowError::Overflow).unwrap();
        self.offsets.push(size);
        match &mut self.validity {
            Some(validity) => validity.push(true),
            None => {}
        }
    }
```

```rust
pub fn primitive_to_binary<T: NativeType + lexical_core::ToLexical, O: Offset>(
    from: &PrimitiveArray<T>,
) -> BinaryArray<O> {

    let builder = from.iter().fold(
        MutableBinaryArray::<O>::with_capacity(from.len()),
        |mut builder, x| {
            match x {
+               Some(x) => unsafe {
+                   builder.reserve(1, T::FORMATTED_SIZE_DECIMAL);
+                   builder.write_values(|bytes| lexical_core::write(*x, bytes).len());
+               },
                None => builder.push_null(),
            }
            builder
```

Improved performance in cast Primitive to Binary/String again (2x) #6

Open sundy-li wants to merge 2 commits into jorgecarleitao:main from datafuse-extras:primitive-cast3

Conversation 3 | Commits 2 | Checks 0 | Files changed 4

sundy-li commented 20 hours ago • edited — Contributor

Memcpy style write, no extra copy.

```
cast int32 to binary 512
                        time:   [7.2389 us 7.2844 us 7.3306 us]
                        change: [-32.305% -31.940% -31.566%] (p = 0.00 < 0.05)
                        Performance has improved.
Found 8 outliers among 100 measurements (8.00%)
```

Arrow2 PR #651

No extra allocation/copy

Case 5: runtime_tracker

```
116 +
117 -     pub fn get_memory_tracker(&self) -> Arc<MemoryTracker> {
118 -         self.memory_tracker.clone()
119       }
120
```

```
160
161 +     #[inline]
162 +     pub fn get_memory_tracker(&self) -> &MemoryTracker {
163 +         &self.memory_tracker
164       }
165
```

```
56 +
57 +     #[inline]
58 +     pub fn alloc_memory(size: i64) {
59 +         unsafe {
60 +             if !TRACKER.is_null() {
61 +                 (*TRACKER).untracked_memory += size;
62 +
63 +                 if (*TRACKER).untracked_memory > UNTRACKED_MEMORY_LIMIT {
64 +                     (*TRACKER)
65 +                         .rt_tracker
66 +                         .memory_tracker
67 +                         .alloc_memory((*TRACKER).untracked_memory);
68 +                     (*TRACKER).untracked_memory = 0;
69 +                 }
70 +             }
71 +         }
72 +     }
73 +
```

Databend PR #3140

Less extra function call

Case 6: extra function call in closure

3 src/compute/sort/primitive/indices.rs

View full | Viewed

```diff
@@ -18,10 +18,11 @@ where
     T: NativeType,
     F: Fn(&T, &T) -> std::cmp::Ordering,
 {
+    let values = array.values().as_slice();
     unsafe {
         common::indices_sorted_unstable_by(
             array.validity(),
-            |x: usize| *array.values().as_slice().get_unchecked(x),
+            |x: usize| *values.get_unchecked(x),
             cmp,
             array.len(),
             options,
```

sundy-li commented on 2 Aug — Contributor — Author

`as_slice` is the bottleneck.

I think we can use `values: &[T]` to replace `get: G`.

Improved in commit: datafuse-extras#1

Now:

```
arrow2-sort-limit 2^13 f32
  time:   [46.601 us 46.665 us 46.728 us]
  change: [-24.847% -24.656% -24.483%] (p = 0.00 < 0.05)
  Performance has improved.
```

Another bottleneck is `from_usize().unwrap()` in the loop.

2

Arrow2 ISSUE #245

Less branch prediction miss

Case 7: unwrap inside loop

```rust
let mut indices: MutableBuffer<I> = unsafe {
    MutableBuffer::from_trusted_len_iter_unchecked(
        iterator: (0..length).map(|x: usize| I::from_usize(index: x).unwrap()),
    )
};

// Soundness:
// indices are by construction `< values.len()`
// limit is by construction `< values.len()`
sort_unstable_by(&mut indices, get, cmp, descending, limit);
```

```rust
} else {
    let mut indices: MutableBuffer<I> = MutableBuffer::from_trusted_len_iter(iterator: I::range(start: 0, end: length).unwrap());

    // Soundness:
    // indices are by construction `< values.len()`
    // limit is by construction `< values.len()`
    sort_unstable_by(&mut indices, get, cmp, descending, limit);
    indices.truncate(len: limit);
    indices.shrink_to_fit();
```

Arrow2 Codes

Less branch prediction miss

Case 8: transform primitive array

```
let result: DFUInt8Array = array.i8()?.iter().map(|x| abs(*x) ).collect();
```

Arrow's format

You, 6 days ago | 1 author (You)

```
#### Column Iteration and validity bitmap combination
To iterate the column, we can use `column.iter()` to generate an iterator, the item is `Option<T>`, `None` means it's null.
But this's inefficient because we need to check the null value every time we iterate the column inside the loop which will pollute the CPU cache.

According to the memory layout of Arrow, we can directly use the validity bitmap of the original column to indicate the null value.
So we have `ArrayApply` trait to help you iterate the column. If there are two zip iterator, we can use `binary` function to combine the validity bitmap of two columns.

ArrayApply
```rust
let array: DFUInt8Array = self.apply_cast_numeric(|a| {
                                AsPrimitive::<u8>::as_(a - (a / rhs) * rhs)
                            });
```
You, 6 days ago • Update docs

binary
```rust
binary(x_series.f64()?, y_series.f64()?, |x, y| x.pow(y))
```
```

Less branch prediction miss

Case 9: null & if combinator in sum

```
select sumIf(number, number > 3) from numbers_mt(N)
```

```
/// Merge the 2 sets of flags (null and if) into a single one. This allows us to use parallelizable sums when
available
const auto * if_flags = assert_cast<const ColumnUInt8&>(*columns[if_argument_pos]).getData().data();
```

```
size_t countBytesInFilterWithNull(const IColumn::Filter & filt, const UInt8 * null_map)
{
    size_t count = 0;

    /** NOTE: In theory, `filt` should only contain zeros and ones.
     * But, just in case, here the condition > 0 (to signed bytes) is used.
     * It would be better to use != 0, then this does not allow SSE2.
     */

    const Int8 * pos = reinterpret_cast<const Int8 *>(filt.data());
    const Int8 * pos2 = reinterpret_cast<const Int8 *>(null_map);
    const Int8 * end = pos + filt.size();

#if defined(__SSE2__) && defined(__POPCNT__)
    const Int8 * end64 = pos + filt.size() / 64 * 64;

    for (; pos < end64; pos += 64, pos2 += 64)
        count += __builtin_popcountll(toBits64(pos) & ~toBits64(pos2));

        /// TODO Add duff device for tail?
#endif

    for (; pos < end; ++pos, ++pos2)
        count += (*pos & ~*pos2) != 0;

    return count;      Alexey Milovidov, 5 years ago • Changed tabs to spaces in code [#CLICKHOUSE-3].
}
```

ClickHouse codes

java - Why is processing a sorted array faster than processing an unsorted array? - Stack Overflow

SIMD

Case 10: simd sum

图 1 ARM7 单周期指令最佳流水线

```
double CommonAdd(double *data, int count)
{
    double result = 0;

    for (int i = 0 ; i < count; i++) {
        result += data[i];
    }

    return result;
}
```

normal version

```
double AVXAdd(double *data, int count)
{
    int offset = 0;

    __m256d v1;
    __m256d sum = _mm256_setzero_pd();

    double ret = 0;

    for (int i = 0; i < count/4; i++) {
        v1 = _mm256_load_pd(data + offset);
        sum = _mm256_add_pd(sum, v1);
        offset += 4;
    }

    sum = _mm256_hadd_pd(sum, sum);

    ret += sum[0];
    ret += sum[2];

    return ret;
}
```

simd version

SIMD in rust

Case 11: null sum in arrow2

```rust
#[multiversion]
#[clone(target = "x86_64+avx")]
fn null_sum_impl<T, I>(values: &[T], mut validity_masks: I) -> T
where
    T: NativeType + Simd,
    T::Simd: Add<Output = T::Simd> + Sum<T>,
    I: BitChunkIterExact<<<T as Simd>::Simd as NativeSimd>::Chunk>,
{
    let mut chunks = values.chunks_exact(T::Simd::LANES);

    let sum = chunks.by_ref().zip(validity_masks.by_ref()).fold(
        T::Simd::default(),
        |acc, (chunk, validity_chunk)| {
            let chunk = T::Simd::from_chunk(chunk);
            let mask = <T::Simd as NativeSimd>::Mask::from_chunk(validity_chunk);
            let selected = chunk.select(mask, T::Simd::default());
            acc + selected
        },
    );

    let remainder = T::Simd::from_incomplete_chunk(chunks.remainder(), T::default());
    let mask = <T::Simd as NativeSimd>::Mask::from_chunk(validity_masks.remainder());
    let remainder = remainder.select(mask, T::Simd::default());
    let reduced = sum + remainder;

    reduced.simd_sum()
}
```

Jorge Leitao, 7 months ago • Refactored aggregate. (#54)

rust-lang/portable-simd: The testing ground for the future of portable SIMD in Rust (github.com)

packed_simd_2 - Rust (rust-lang.github.io)

Auto vectorized


```
arrow2-sum 2^13 u32     time:   [545.01 ns 546.26 ns 547.72 ns]
                        change: [-0.2234% +0.2358% +0.8664%] (p = 0.41 > 0.05)
                        No change in performance detected.

arrow2-sum 2^13 u32 nosimd
                        time:   [316.59 ns 317.36 ns 318.20 ns]
                        change: [+0.4290% +0.8618% +1.2727%] (p = 0.00 < 0.05)
                        Change within noise threshold.
arrow2-sum null 2^13 u32
                        time:   [4.3197 us 4.3290 us 4.3394 us]
                        change: [-0.1470% +0.3333% +0.8057%] (p = 0.19 > 0.05)

arrow2-sum null 2^13 u32 nosimd
                        time:   [10.149 us 10.168 us 10.190 us]
                        change: [-0.8429% -0.4038% +0.1192%] (p = 0.11 > 0.05)
```

```rust
fn nosimd_sum_u32(arr_a: &PrimitiveArray<u32>) -> Option<u32> {
    let mut sum = 0;
    if arr_a.null_count() == 0 {
        arr_a.values().as_slice().iter().for_each(|f| {
            sum += *f;
        });
    } else {
        if let Some(c) = arr_a.validity() {
            arr_a
                .values()
                .as_slice()
                .iter()
                .zip(c.iter())
                .for_each(|(f, is_null)| sum += *f * (is_null as u32));
        }
    }
```

[Auto vectorized vs packed_simd in arrow · Issue #255 · jorgecarleitao/arrow2 (github.com)]

Aggregation optimization

A Long and sad story to tell

```
SELECT max(number), sum(number) FROM numbers_mt(1000000000)
GROUP BY number % 3, number % 4, number % 5 LIMIT 10;
```

This SQL works extremely slow in databend’s early versions.

Group by transform

Early version:
split the blocks by hash key

It's really poor idea!!!

Group by transform

Never split the blocks

Varint hash method


```rust
pub enum HashMethodKind {
    Serializer(HashMethodSerializer),
    KeysU8(HashMethodKeysU8),
    KeysU16(HashMethodKeysU16),
    KeysU32(HashMethodKeysU32),
    KeysU64(HashMethodKeysU64),
}
```

```rust
impl DataBlock {
    pub fn choose_hash_method(
        block: &DataBlock,
        column_names: &[String],
    ) -> Result<HashMethodKind> {
        let mut group_key_len: usize = 0;
        for col: &String in column_names {
            let column: &DataColumn = block.try_column_by_name(col)?;
            let typ: DataType = column.data_type();
            if typ.is_integer() {
                group_key_len += typ.numeric_byte_size()?;
            } else {
                return Ok(HashMethodKind::Serializer(HashMethodSerializer::default()));
            }
        }      sundy-li, 5 months ago • [groupby] add KeysU8/KeysU16/KeysU32/KeysU64 hash…
        match group_key_len {
            1 => Ok(HashMethodKind::KeysU8(HashMethodKeysU8::default())),
            2 => Ok(HashMethodKind::KeysU16(HashMethodKeysU16::default())),
            3..=4 => Ok(HashMethodKind::KeysU32(HashMethodKeysU32::default())),
            5..=8 => Ok(HashMethodKind::KeysU64(HashMethodKeysU64::default())),
            _ => Ok(HashMethodKind::Serializer(HashMethodSerializer::default())),
        }
    }
```

This makes hashmap much smaller and efficient

Memory ARENA

Memory ARENA by using `Bumpalo crate`

States are much more cache friendly.

```
96
97      let arena: Bump = Bump::new();
98
99      let (layout: Layout, offsets_aggregate_states: Vec<usize>) = unsafe { get_layout_offsets(&funcs) };
100
101     let places: Vec<usize> = {
102         let place: StateAddr = arena.alloc_layout(layout).into();
103         funcs: Vec<Arc<dyn AggregateFunction>>
104             .iter(): Iter<Arc<dyn AggregateFunction>>      sundy-li, 4 months ago • aggregator usize
105             .enumerate(): impl Iterator<Item = (usize, …)>
106             .map(|(idx: usize, func: &Arc<dyn AggregateFunction>)| {
107                 let arg_place: StateAddr = place.next(offset: offsets_aggregate_states[idx]);
108                 func.init_state(arg_place);
109                 arg_place.addr()
110             }): impl Iterator<Item = usize>
111             .collect()
112     };
113
```

This makes hashmap much smaller and efficient.

Memory ARENA

Memory ARENA by using `Bumpalo crate`

States are much more cache friendly.

```
96
97      let arena: Bump = Bump::new();
98
99      let (layout: Layout, offsets_aggregate_states: Vec<usize>) = unsafe { get_layout_offsets(&funcs) };
100
101     let places: Vec<usize> = {
102         let place: StateAddr = arena.alloc_layout(layout).into();
103         funcs: Vec<Arc<dyn AggregateFunction>>
104             .iter(): Iter<Arc<dyn AggregateFunction>>      sundy-li, 4 months ago • aggregator usize
105             .enumerate(): impl Iterator<Item = (usize, …)>
106             .map(|(idx: usize, func: &Arc<dyn AggregateFunction>)| {
107                 let arg_place: StateAddr = place.next(offset: offsets_aggregate_states[idx]);
108                 func.init_state(arg_place);
109                 arg_place.addr()
110             }): impl Iterator<Item = usize>
111             .collect()
112     };
113
```

This makes hashmap much smaller and efficient.

States combination


```rust
/// # Safety
/// layout must ensure to be aligned
pub unsafe fn get_layout_offsets(funcs: &[AggregateFunctionRef]) -> (Layout, Vec<usize>) {
    let mut align_aggregate_states: usize = 0;
    let mut total_size_aggregate_states: usize = 0;

    let mut offsets_aggregate_states: Vec<usize> = Vec::with_capacity(funcs.len());
    for i: usize in 0..funcs.len() {
        offsets_aggregate_states.push(total_size_aggregate_states);
        let layout: Layout = funcs[i].state_layout();
        sundy-li, 5 months ago • [functions] support state arena
        total_size_aggregate_states += layout.size();
        align_aggregate_states = align_aggregate_states.max(layout.align());
        // Not the last aggregate_state
        if i + 1 < funcs.len() {
            let next_layout: Layout = funcs[i + 1].state_layout();
            total_size_aggregate_states = (total_size_aggregate_states + next_layout.align() - 1)
                / next_layout.align()
                * next_layout.align();
        }
    }
    let layout: Layout =
        Layout::from_size_align_unchecked(size: total_size_aggregate_states, align_aggregate_states);
    (layout, offsets_aggregate_states)
}
```

States could share a single address with specific layout offset

CPU pipeline

Div & Rem operators are very slow in databend!

number % 3, number / 3

Why?

	Integer			Floating point		
Operation	Latency	Issue	Capacity	Latency	Issue	Capacity
Addition	1	1	4	3	1	1
Multiplication	3	1	1	5	1	2
Division	3–30	3–30	1	3–15	3–15	1

Figure 5.12 Latency, issue time, and capacity characteristics of reference machine operations. Latency indicates the total number of clock cycles required to perform the actual operations, while issue time indicates the minimum number of cycles between two independent operations. The capacity indicates how many of these operations can be issued simultaneously. The times for division depend on the data values.

Improve constant div

Transform mod to multiply operator

```
uint32_t d = ...;// your divisor > 0

uint64_t c = UINT64_C(0xFFFFFFFFFFFFFFFF) / d + 1;

// fastmod computes (n mod d) given precomputed c
uint32_t fastmod(uint32_t n ) {
  uint64_t lowbits = c * n;
  return ((__uint128_t)lowbits * d) >> 64;
 }
```

4X - 10X improvement

Refer to:

Blog:
https://lemire.me/blog/2019/02/08/faster-remainders-when-the-divisor-is-a-constant-beating-compilers-and-libdivide/
Paper: https://arxiv.org/abs/1902.01961
Go: https://github.com/bmkessler/fastdiv
Rust: https://docs.rs/strength_reduce/0.2.3/strength_reduce/

Aggregation optimization

Sad story reaches happy ending

After some more optimizations,
GroupBy performance is as good as ClickHouse now

3 Pipeline/Dataflow executor

Volcano model

火山模型行处理模型

BatchData

Pull vs pushed

【数据库内核】物理计算引擎Push模型之编译执行_Night_ZW的博客-CSDN博客_数据库编译执行

Compile vs vectorized execution


```
query(...)
  // build hash table
  for(i = 0; i < S.size(); i++)
    ht.insert(<S.att1[i], S.att2[i]>, S.att3[i])
  // probe hash table
  for(i = 0; i < R.size(); i++)
    int k1 = R.att1[i]
    string* k2 = R.att2[i]
    int hash = hash(k1, k2)
    for(Entry* e = ht.find(hash); e; e = e->next)
      if(e->key1 == k1 && e->key2 == *k2)
        ... // code of parent operator
```

(a) Code generated for hash join

```
class HashJoin
  Primitives probeHash_, compareKeys_, buildGather_;
  ...
int HashJoin::next()
  ... // consume build side and create hash table
  int n = probe->next()// get tuples from probe side
  // *Interpretation*: compute hashes
  vec<int> hashes = probeHash_.eval(n)
  // find hash candidate matches for hashes
  vec<Entry*> candidates = ht.findCandidates(hashes)
  // matches: int references a position in hashes
  vec<Entry*, int> matches = {}
  // check candidates to find matches
  while(candidates.size() > 0)
    // *Interpretation*
    vec<bool> isEqual = compareKeys_.eval(n, candidates)
    hits, candidates = extractHits(isEqual, candidates)
    matches += hits
  // *Interpretation*: gather from hash table into
  // buffers for next operator
  buildGather_.eval(matches)
  return matches.size()
```

(b) Vectorized code that performs a hash join

http://cidrdb.org/cidr2005/papers/P19.pdf

https://www.vldb.org/pvldb/vol11/p2209-kersten.pdf

Compile vs vectorized execution


```
query(...)
  // build hash table
  for(i = 0; i < S.size(); i++)
    ht.insert(<S.att1[i], S.att2[i]>, S.att3[i])
  // probe hash table
  for(i = 0; i < R.size(); i++)
    int k1 = R.att1[i]
    string* k2 = R.att2[i]
    int hash = hash(k1, k2)
    for(Entry* e = ht.find(hash); e; e = e->next)
      if(e->key1 == k1 && e->key2 == *k2)
        ... // code of parent operator
```

(a) Code generated for hash join

```
class HashJoin
  Primitives probeHash_, compareKeys_, buildGather_;
  ...
int HashJoin::next()
  ... // consume build side and create hash table
  int n = probe->next()// get tuples from probe side
  // *Interpretation*: compute hashes
  vec<int> hashes = probeHash_.eval(n)
  // find hash candidate matches for hashes
  vec<Entry*> candidates = ht.findCandidates(hashes)
  // matches: int references a position in hashes
  vec<Entry*, int> matches = {}
  // check candidates to find matches
  while(candidates.size() > 0)
    // *Interpretation*
    vec<bool> isEqual = compareKeys_.eval(n, candidates)
    hits, candidates = extractHits(isEqual, candidates)
    matches += hits
  // *Interpretation*: gather from hash table into
  // buffers for next operator
  buildGather_.eval(matches)
  return matches.size()
```

(b) Vectorized code that performs a hash join

http://cidrdb.org/cidr2005/papers/P19.pdf

https://www.vldb.org/pvldb/vol11/p2209-kersten.pdf

Compile execution


```
select   *
from     R1,R3,
         (select   R2.z,count(*)
         from      R2
         where     R2.y=3
         group by  R2.z) R2
where    R1.x=7 and R1.a=R3.b and R2.z=R3.c
```

Figure 2: Example Query

Figure 3: Example Execution Plan for Figure 2

initialize memory of $\bowtie_{a=b}$, $\bowtie_{c=z}$, and Γ_z
for each tuple t in R_1
  if $t.x = 7$
    materialize t in hash table of $\bowtie_{a=b}$
for each tuple t in R_2
  if $t.y = 3$
    aggregate t in hash table of Γ_z
for each tuple t in Γ_z
  materialize t in hash table of $\bowtie_{z=c}$
for each tuple t_3 in R_3
  for each match t_2 in $\bowtie_{z=c}[t_3.c]$
    for each match t_1 in $\bowtie_{a=b}[t_3.b]$
      output $t_1 \circ t_2 \circ t_3$

Figure 4: Compiled query for Figure 3

https://www.vldb.org/pvldb/vol4/p539-neumann.pdf

Databend

THANK YOU FOR WATCHING