GridGain

Best Practices For Loading Data To Distributed Systems With Change Data Capture

Alexey Goncharuk

Change Data Capture In The Wild

Agenda

- What is CDC?
- What can I do with CDC?
- What is available in Ignite / GridGain?

GridGain

What is Change Data Capture?

What is Change Data Capture

- Have a data set or arbitrary size
- Determine what records changed since a given moment
- Many ways to achieve this...

What Is CDC?

Record Change Markers

- Timestamps
- Versions
- Statuses
- Attached to application data model

Record Change Markers

ID	…	UPDATE_TS
1		2019-10-10 00:01:02.000
2		2019-10-09 11:01:02.000
3		2018-10-09 18:36:13.000
4		2019-09-01 01:02:03.000
…		
10		2019-06-13 11:12:04.000

Record Change Markers

ID	…	UPDATE_TS
1		2019-10-10 00:01:02.000
2		2019-10-09 11:01:02.000
3		2018-10-09 18:36:13.000
4		**2019-11-01 23:59:59.000**
…		
10		**2019-11-15 14:00:00.000**

Record Change Markers

ID	…	UPDATE_TS
1		2019-10-10 00:01:02.000
2		2019-10-09 11:01:02.000
3		2018-10-09 18:36:13.000
4		**2019-11-01 23:59:59.000**
…		
10		**2019-11-15 14:00:00.000**

SELECT * FROM Table WHERE UPDATE_TS > ’ 2019-11-01 00:00:00.000’

GridGain

Record Change Markers

Cons

- Detecting changes is tricky
  - Full scan
  - Additional index for change markers
- No previous value (change coalescing)

Record Change Markers

Pros

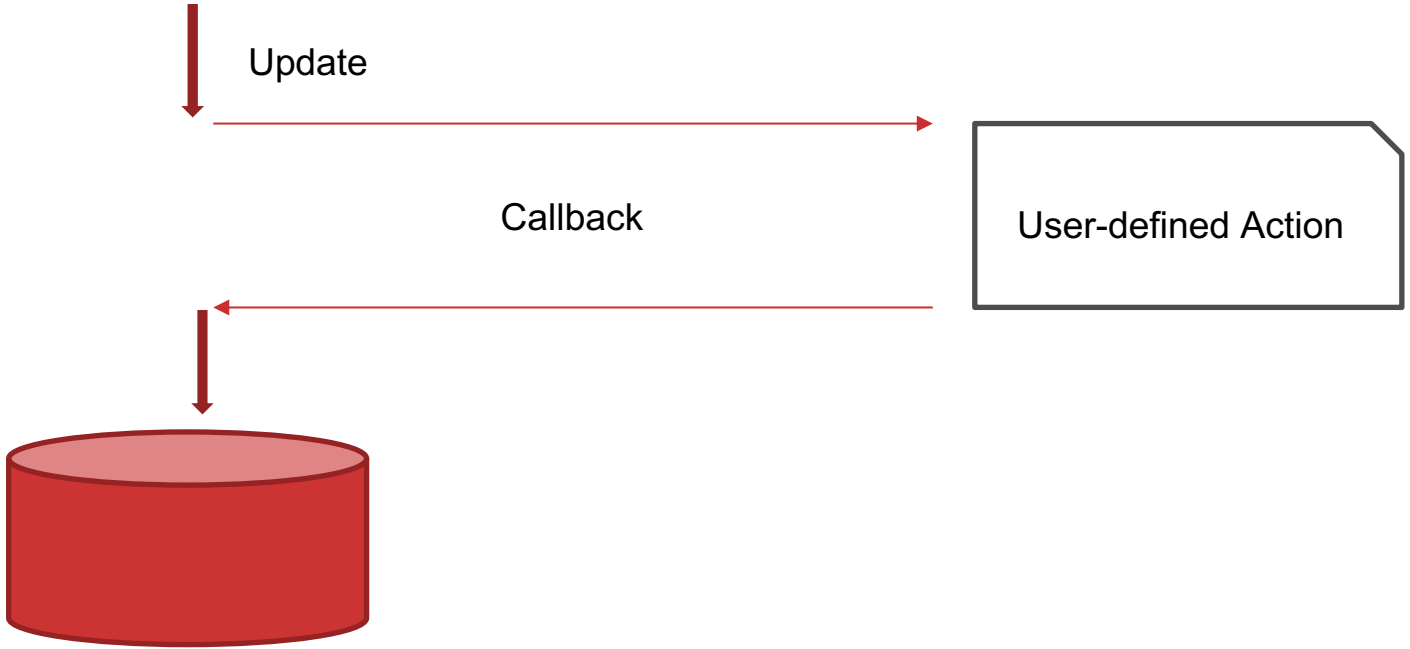
- May be implemented in application layer
- Delayed change consumption
- Negligible storage overhead

What Is CDC?

Callbacks

- Triggers / interceptors / etc...
- User code is supplied to the storage system

Callbacks



Cons

- Invoked synchronously
- Tricky failover in distributed systems

Callbacks

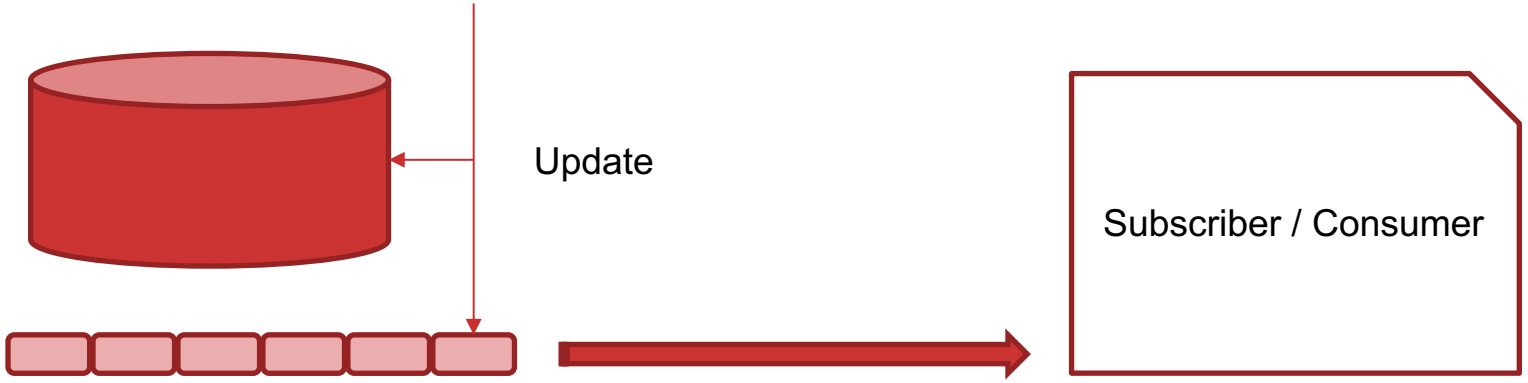
Pros

- No system storage/insert overhead
- Previous value is usually available
- May have an ability to modify updated value

Change Feed

- Changes are stored as events (Event Sourcing)
- Or changes produce events
- Consumers subscribe to a change feed
- Database WAL is an events source!

Change Feed



Change Feed

Cons

- Need additional storage to keep changes

Change Feed

Pros

- Previous values are usually available
- Full change history is preserved
- Possibly an ability to re-read the history

CDC Applications

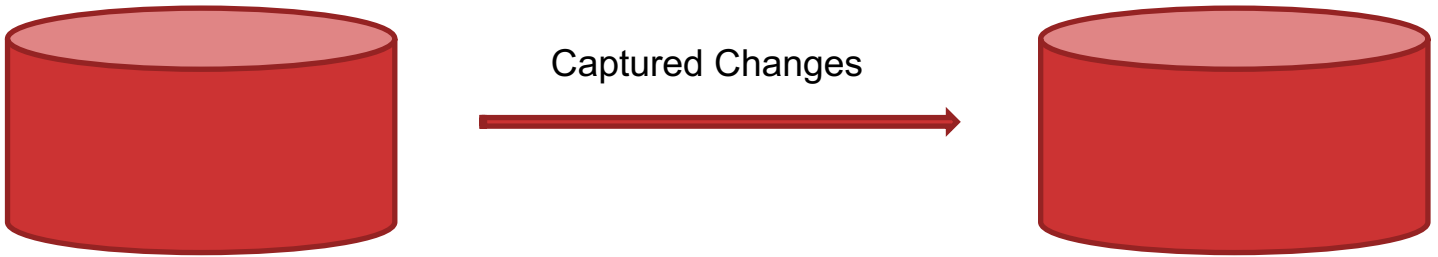
GridGain

CDC Applications

Continuous Data Integration

- “Active” database produces changes
- The changes are applied to a secondary system

Continuous Data Integration



CDC Applications

Continuous Data Integration

- Reads offload
- Audit Changelog
- Cross-system Replication
- High Availability

Running function calculation

- Computationally expensive function over a large set of items?
- Calculate once, then apply deltas

Running function calculation

- AVG (ITEMS) = SUM (ITEMS) / COUNT (ITEMS)
  - O(N) Complexity
- On insert => SUM += New Value, COUNT += 1
- On delete => SUM -= Deleted Value, COUNT -= 1
- On update => SUM = SUM - Old Value + New Value
- Average is a O(1) operation

Cross-System Active-Active Replication

- Updates feed is going both ways
- Need to resolve conflicts
- Conflict-free Replicated Data Types (CRDTs) for help

Basic CRDTs

- Grow-only counter
- Positive-negative counter
- Grow-only set
- Two-phase set
- Last-write-wins
- …

CDC In Apache Ignite

GridGain

Applying Changes To Ignite

- `IgniteDataStreamer` to optimally deliver changes to data nodes
- A user can use custom stream receiver
- Out-of-the-box integrations
  - Kafka
  - MQTT
  - …

Callbacks

- `CacheInterceptor`
  - Guarantees update order
  - May alter inserted value
  - Synchronous, may affect performance

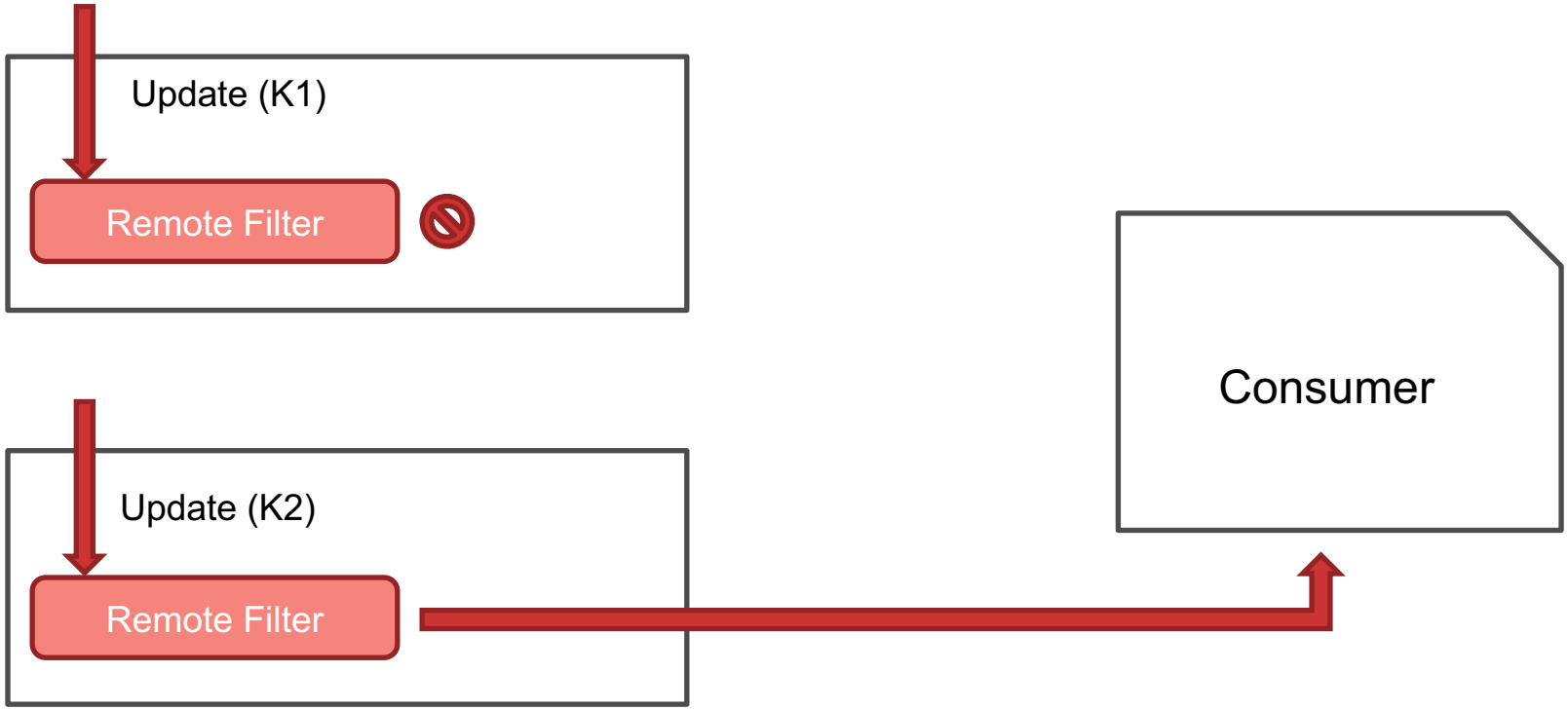
Callbacks

- Cache Events
  - Guarantee update order
  - Asynchronous

Callbacks And Change Feed Combined

- `ContinuousQuery`
  - Client - server subscription
  - Remote filter acts as a synchronous callback
  - Local listener acts as a sink

CDC In Ignite



Callbacks And Change Feed Combined

- Automatic failover in case of primary node crash
- Single-key ordering guarantees

CDC In Ignite

- Ingestion
  - `IgniteDataStreamer`
- Capturing Changes
  - `CacheInterceptor`
  - Events
  - `ContinuousQuery`

Summary

- CDC is a powerful and a well-known technique
- Many systems have built-in support for CDC
- May improve both development time and performance

Want To Contribute?

- dev@ignite.apache.org
- agoncharuk@apache.org

Q&A

Thank you for your attention!

GridGain