KIP-870: Retention policy based on record event time

Status

Current state: "Under Discussion"

Discussion thread: *here*

JIRA: KAFKA-13866

Motivation

Time-based retention policy compares the record timestamp to broker wall-clock time. Those semantics are questionable and also lead to issues for data reprocessing: If one want to re-process older data then retention time, it's not possible as broker expire those record aggressively and user need to increate the retention time accordingly.

Especially for Kafka Stream, we have seen many cases when users got bit by the current behavior.

It would be best, if Kafka would track *two* timestamps per record: the record event-time (as the broker do currently), plus the log append-time (which is only tracked currently if the topic is configured with "append-time" tracking, but the issue is, that it overwrite the producer provided record event-time).

Tracking both timestamps would allow to set a pure wall-clock time retention time plus a pure event-time retention time policy:

- Wall-clock time: keep (at least) the date X days after writing
- Event-time: keep (at max) the X days worth of event-time data

Public Interfaces

Binary log format

Both, broker wall-clock and event timestamp will be stored in message.

Partition data

Maximum value of record event time written in partition must be tracked.

Configuration

New configuration option to setup record event time retention policy:

- `retention.max.eventtime.ms` - Maximum amount of milliseconds to keep data based on event-time.

If user will try to set `retention.max.eventtime.ms`

Proposed Changes

- Modify binary log format to store both - event broker time and record event time.
- Track per partition max event time value - `partition.maxEventTime.`
- If `retention.max.eventtime.ms` unset - topic truncation works as before the changes.
- If `retention.max.eventtime.ms` set - decision of truncation made from the combining previously existing conditions and the following clause -

clause.java

```
partition.maxEventTime - record.recordTime > config.get("retention.max.ms.eventtime")
```

Compatibility, Deprecation, and Migration Plan

- What impact (if any) will there be on existing users?

User will be able to configure retention policy based on record event time for newly created topics.

Test Plan

Test must be done for all properties combination and for topics that was created after and before changes.

Rejected Alternatives

User can increase retention.ms to make sure data required for reprocessing not deleted by the broker.