KIP-644: Support sources which can't give out changed data in kafka connect

Status

Current state: *Under Discussion*

Discussion thread: -

JIRA: -

Please keep the discussion on the mailing list rather than commenting on the wiki (wiki discussions get unwieldy fast).

Motivation

There are source which always give out the entire data when polled, there is no way to query these sources for the changed data. Motivation is to support such sources in kafka connect so that kafka connect can publish the incremental (changed) data for such source.

Public Interfaces

1. The contract of the poll method in SourceTask will be modified to additionally return a list of state records.
2. SourceTaskContext interface will now support a new method to get state storage reader.
3. An internal topic will be introduced to store the state of the source.

Proposed Changes

For such type of sources, it is required to calculate the changed data (data which have changed since the last poll) within the kafka connect. Kafka connect can support a way to allow such sources to maintain the source state, every time the source is polled the current source state is used to compute the changed data which is published to kafka. The source state will get updated after every poll. The kafka connect also needs to periodically persist the state for recovery post restarts/crash.

Compatibility, Deprecation, and Migration Plan

- To be planned out

Rejected Alternatives

- None