

Streaming ETL

With Flink and Elasticsearch

Jared Stehler | @jstehler

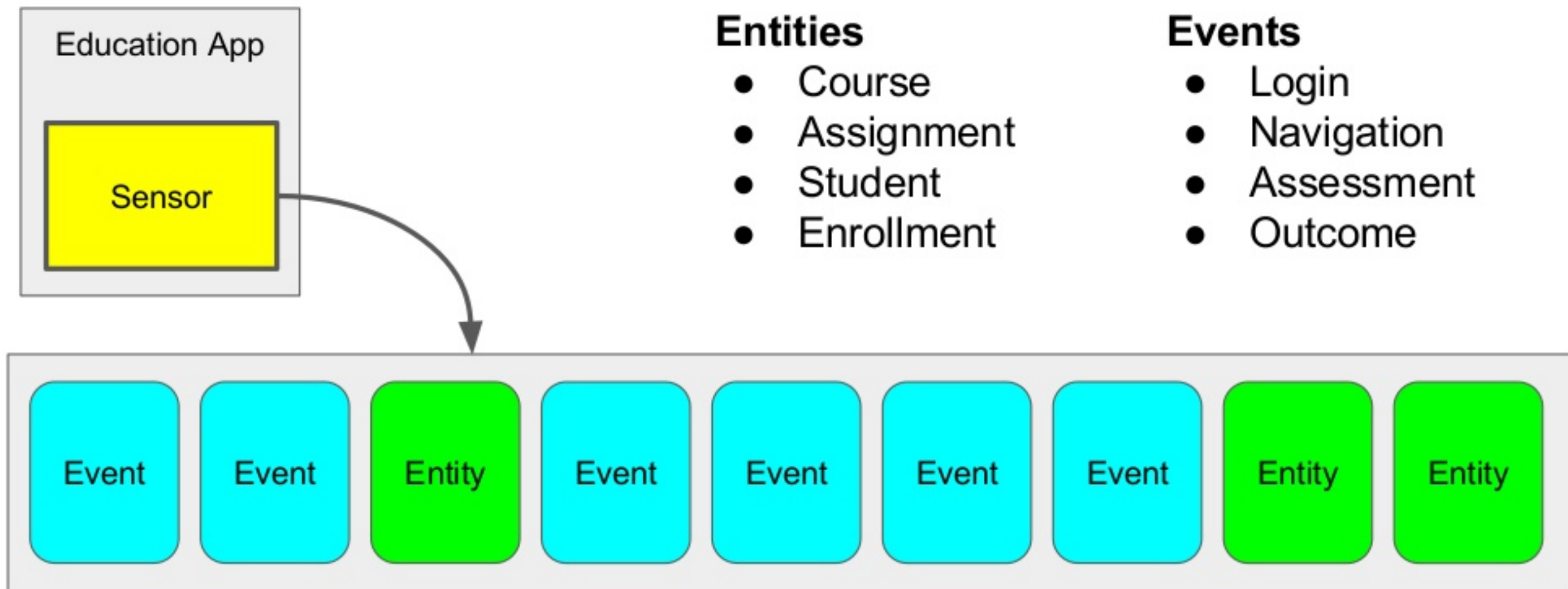
About Me

- Principal Architect @ [Edgenuity](#) - Online Learning Software for K-12
 - Previously @ Intellify Learning
- I Live in Boston, MA
- I Work on Analytics Systems for Educational (EdTech) Software

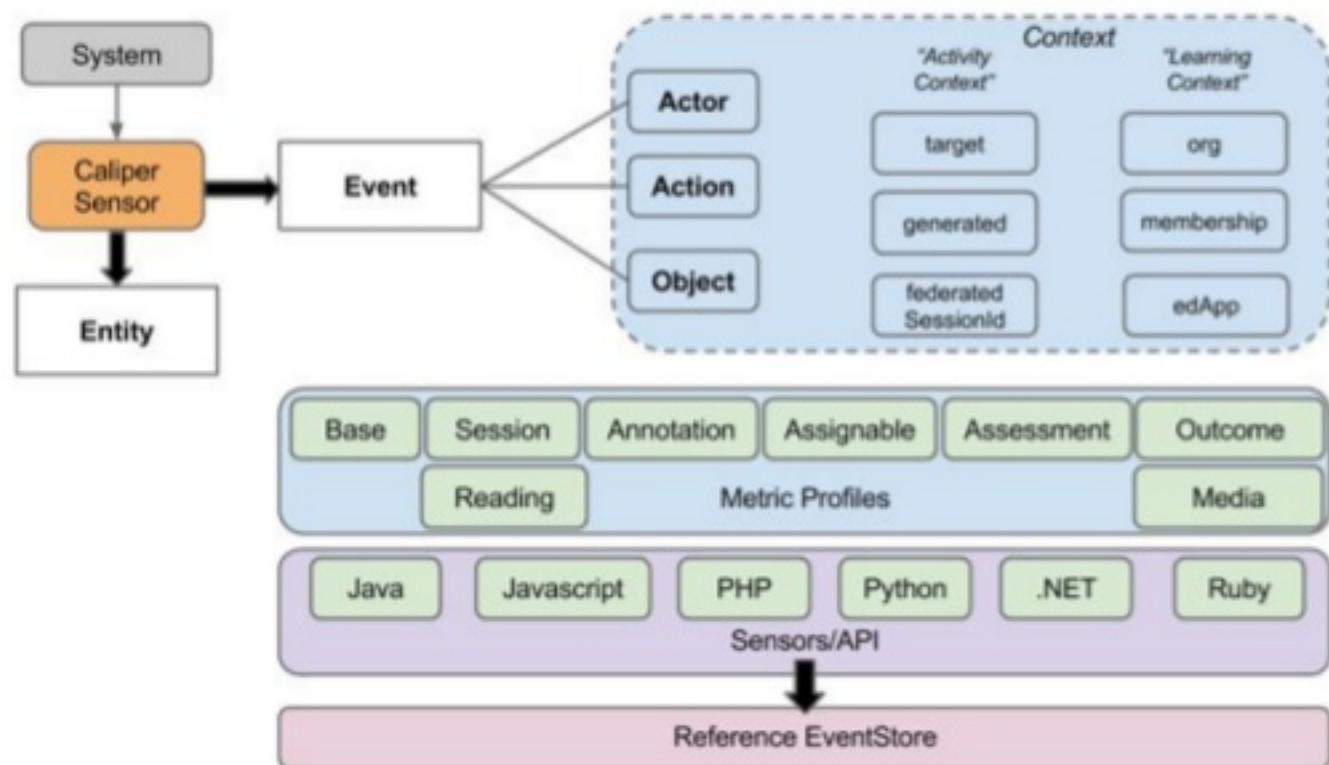
Agenda

- Background
- ETL Approach
- Data Pipeline
- Building, Deploying, Running

Source Data



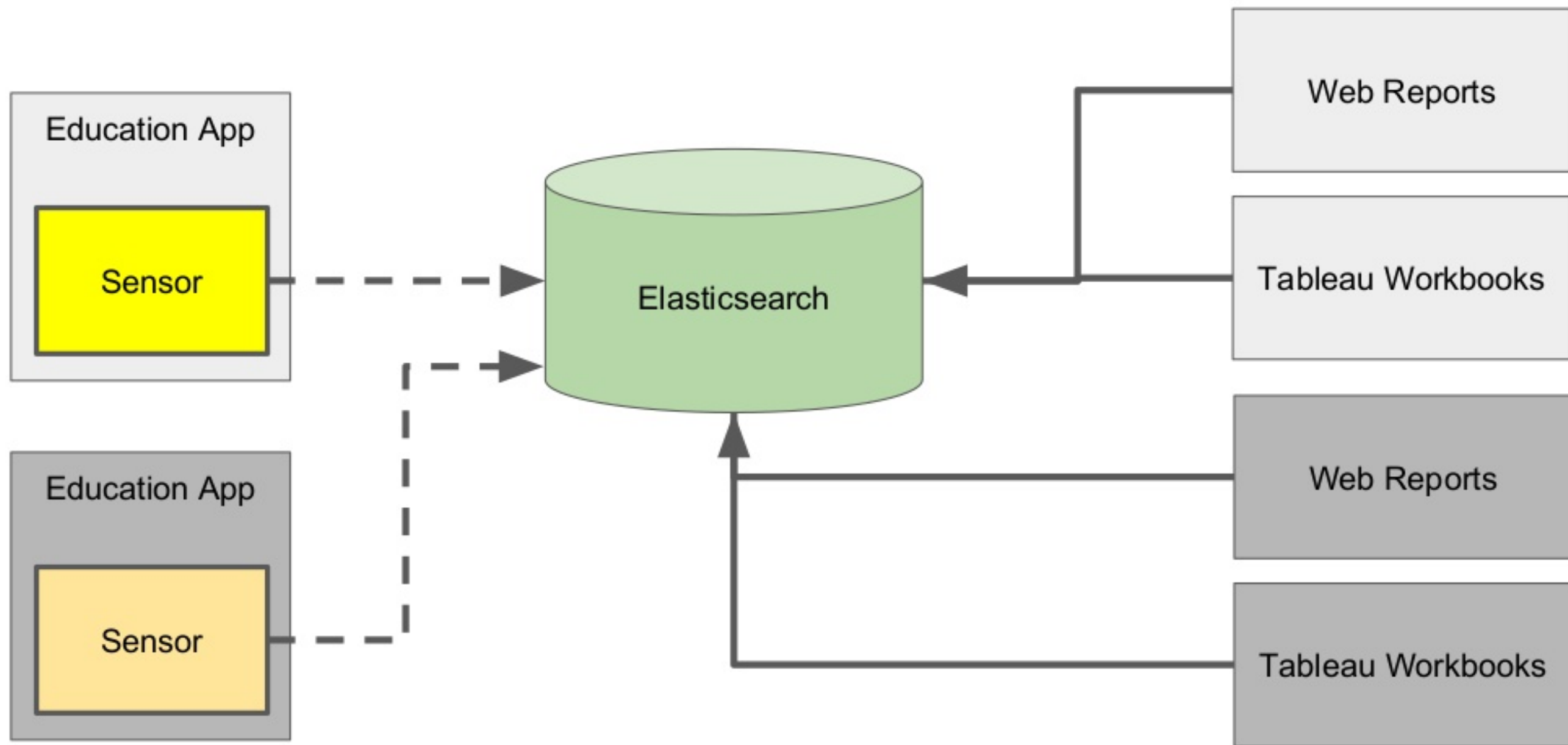
Caliper Analytics Specification



- Entities
 - name
 - dateModified
 - Extensions
- Events
 - Shared Attributes
 - eventTime
 - Actor
 - Action
 - Object
 - Specific Types
 - Session
 - Annotation
 - Navigation
 - Assessment

Example Engagement Sequence

Actors	Student, LMS, Reader App, Quiz App
Basic Flow of Events	<ol style="list-style-type: none"> 1. Student logs in to interact with a course. A SessionEvent is sent by the sensor with an action of logged in. 2. The student navigates to page in the reading. A NavigationEvent is sent by the sensor with an action of navigated to. 3. The student adds a tag to the page. A TagAnnotation is generated. An AnnotationEvent is sent by the sensor with an action of tagged.. 4. The student starts a quiz and generates an Attempt. An AssessmentEvent is sent by the sensor with an action of started.. 5. The student starts question 1. Generates an Attempt. An AssessmentItemEvent is sent by the sensor with an action of started. 6. The student completes question 1 and generates a response. An AssessmentItemEvent is sent by the sensor with an action of completed. 7. Repeats 5-6 for all questions 8. The student submits the quiz. An AssessmentEvent is sent by the sensor with an action of submitted. 9. System grades the quiz and generates a Result An OutcomeEvent is sent by the sensor with an action of graded 10. The student logs out of the course. A SessionEvent is sent by the sensor with an action of logged out.



Problems with Initial Approach

- Flexible Input Schema, Non-Uniform Input Data
- Everything in One Index - Query Performance Problems
- Elasticsearch Doesn't Do Joins!

ETL Solution

- Unified Set of Table Schemas - Avro
- Flink Job per Input Source - Map / Transform to Avro Schema
- Store Outputs in Index per Table - Elasticsearch
- Store Outputs in S3 as Parquet

ETL Join Example

{event} - attempt

- generated.actor.@id
- generated.score
- object.assignable.@id

<<avro>>

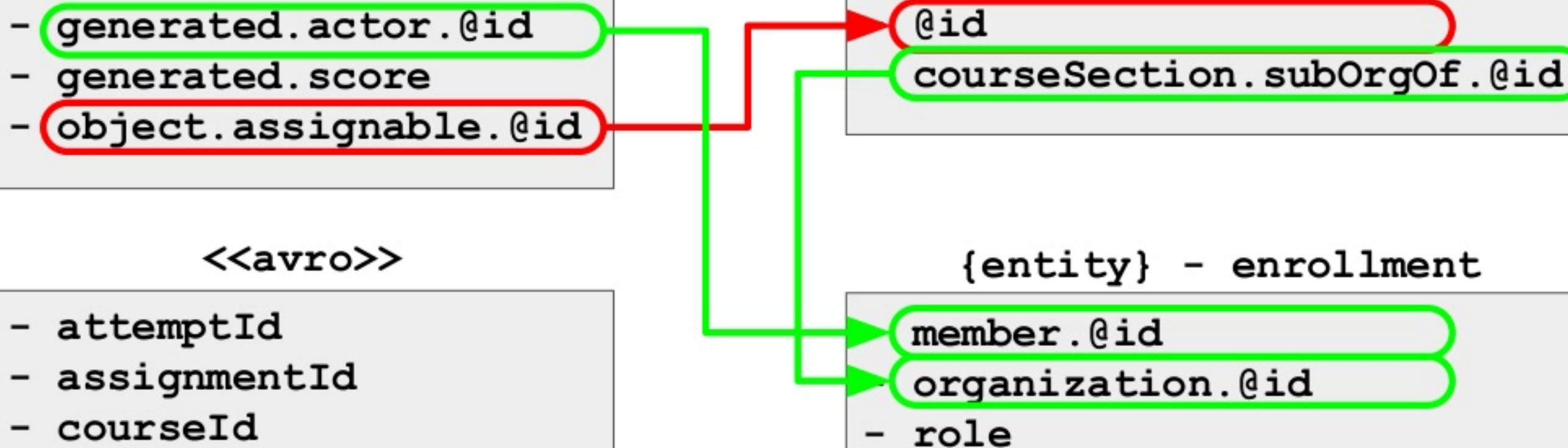
- attemptId
- assignmentId
- courseId
- role
- score

{entity} - assignment

- @id
- courseSection.subOrgOf.@id

{entity} - enrollment

- member.@id
- organization.@id
- role



ETL Join Example - Attempt Event Record

```
{
  "event": {
    "@type": "http://purl.imsglobal.org/caliper/v1/OutcomeEvent",
    "generated": {
      "actor": {
        "@id": "https://example.edu/users/554433"
      },
      "totalScore": 92.4
    },
    "object": {
      "assignable": {
        "@id": "https://example.edu/terms/201801/courses/7/sections/1/assign/2"
      }
    }
  }
}
```

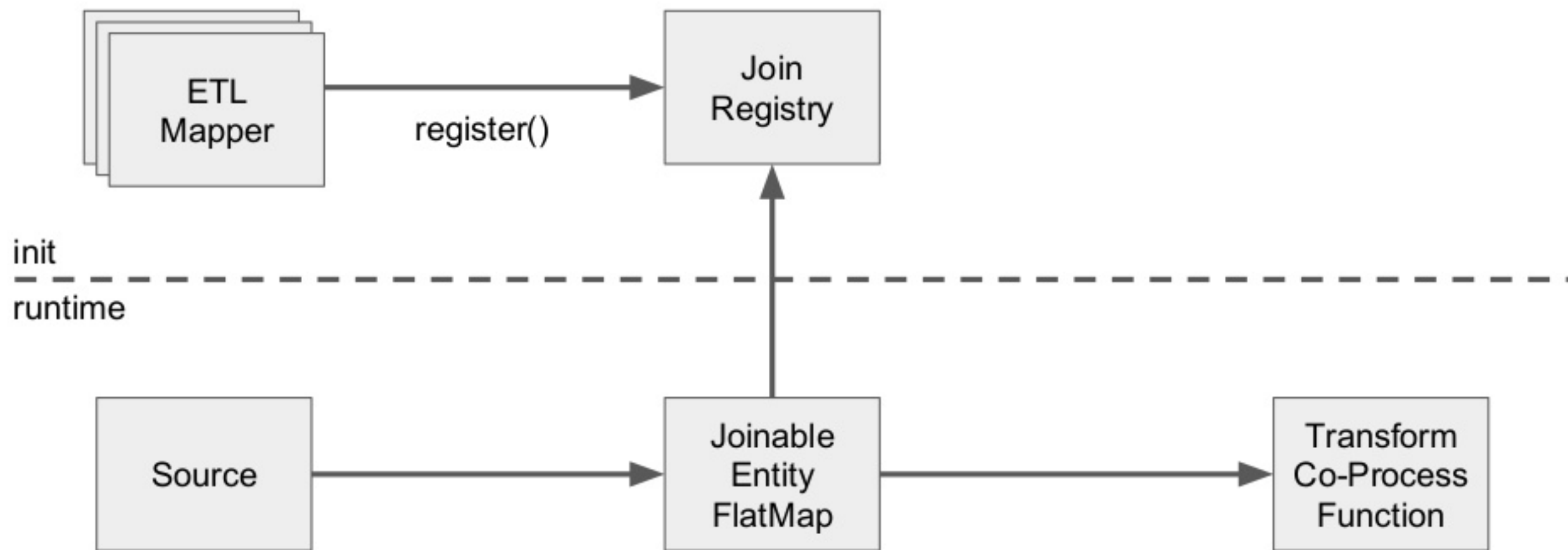
ETL Join Example - Assignment Entity Record

```
{
  "entity": {
    "@type": "http://purl.imsglobal.org/caliper/v1/AssignableDigitalResource",
    "@id": "https://example.edu/terms/201801/courses/7/sections/1/assign/2",
    "extensions": {
      "courseSection": {
        "@id": "https://example.edu/terms/201801/courses/7/sections/1",
        "subOrganizationOf": {
          "@id": "https://example.edu/terms/201801/courses/7"
        }
      }
    }
  }
}
```

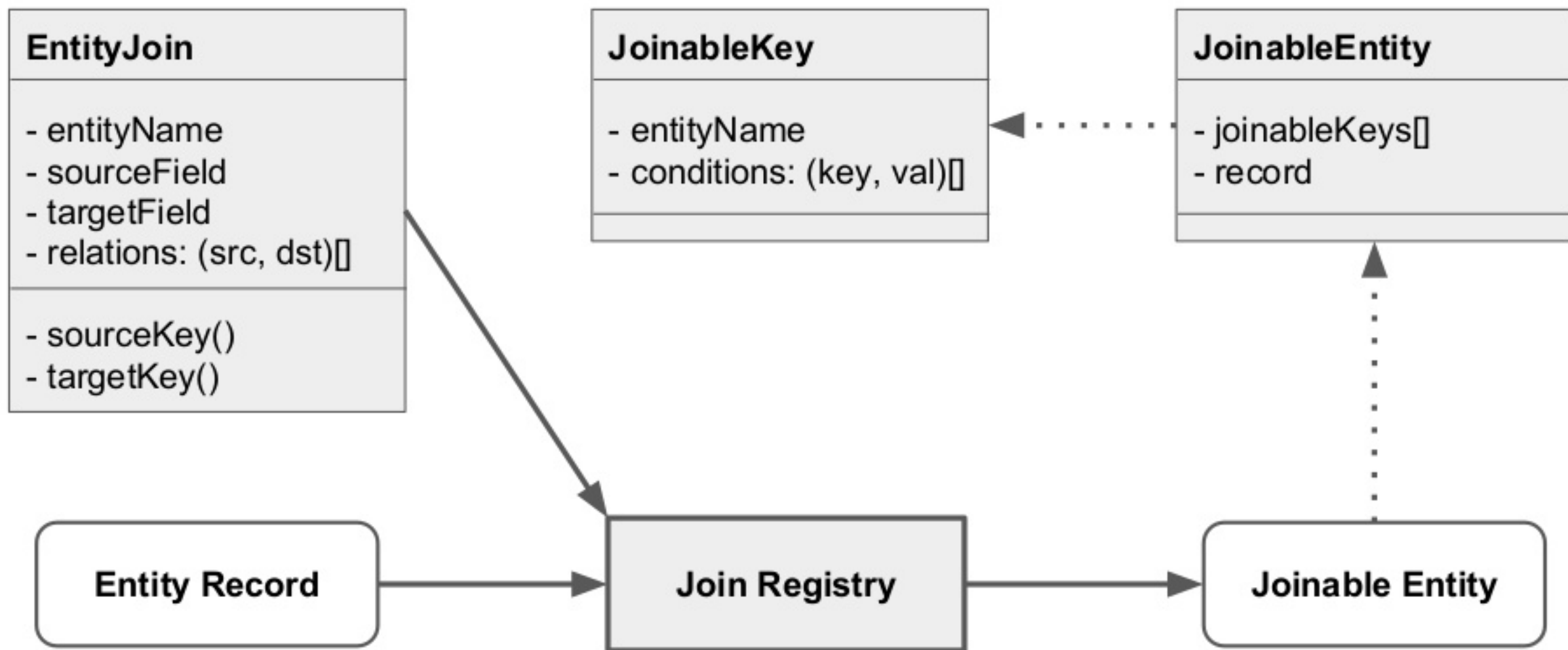
ETL Join Example - Enrollment Entity Record

```
{
  "entity": {
    "@type": "http://purl.imsglobal.org/caliper/v1/lis/Membership",
    "member": {
      "@id": "https://example.edu/users/554433",
      "type": "Person"
    },
    "organization": {
      "@id": "https://example.edu/terms/201801/courses/7/sections/1",
      "type": "CourseSection",
      "subOrganizationOf": {
        "@id": "https://example.edu/terms/201801/courses/7",
        "type": "CourseOffering"
      }
    },
    "roles": [ "Learner" ]
  }
}
```

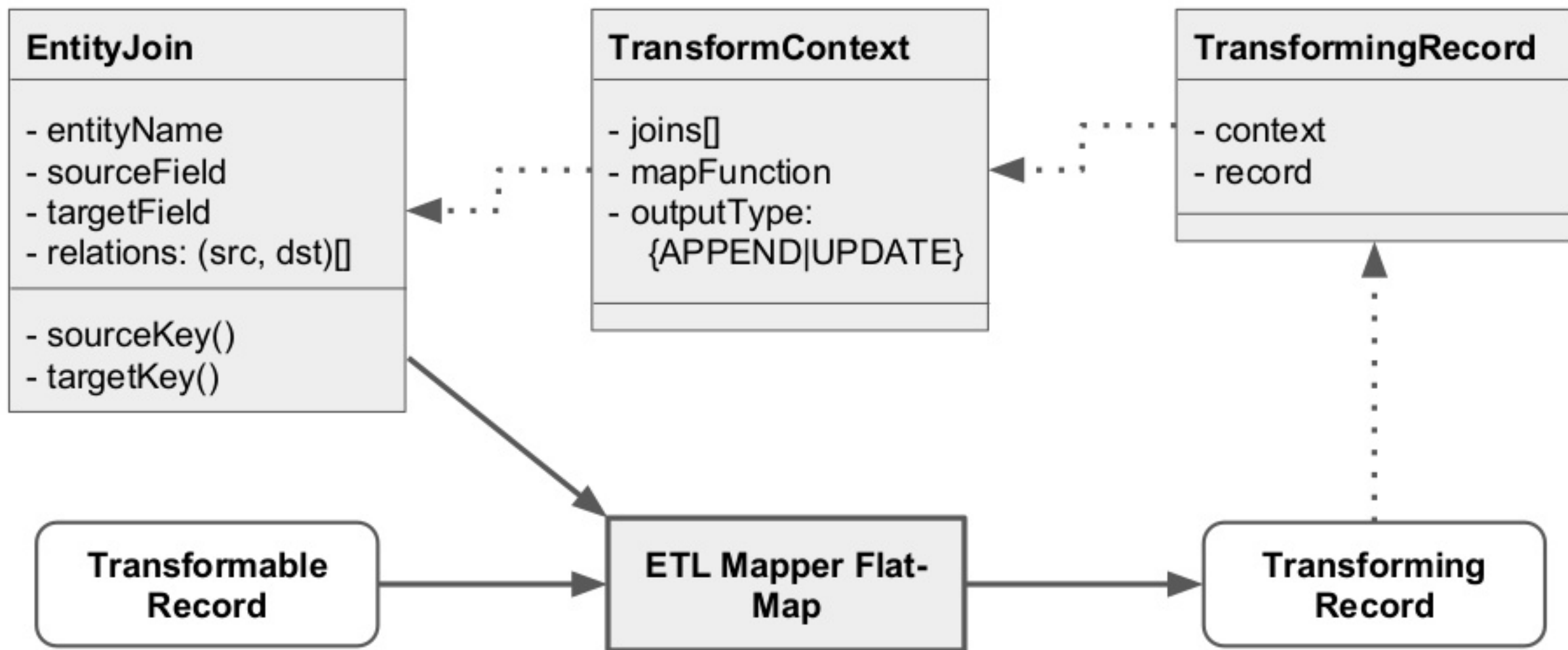
Join Registry



Join State



Transform State



ETL Transforming Record Lifecycle

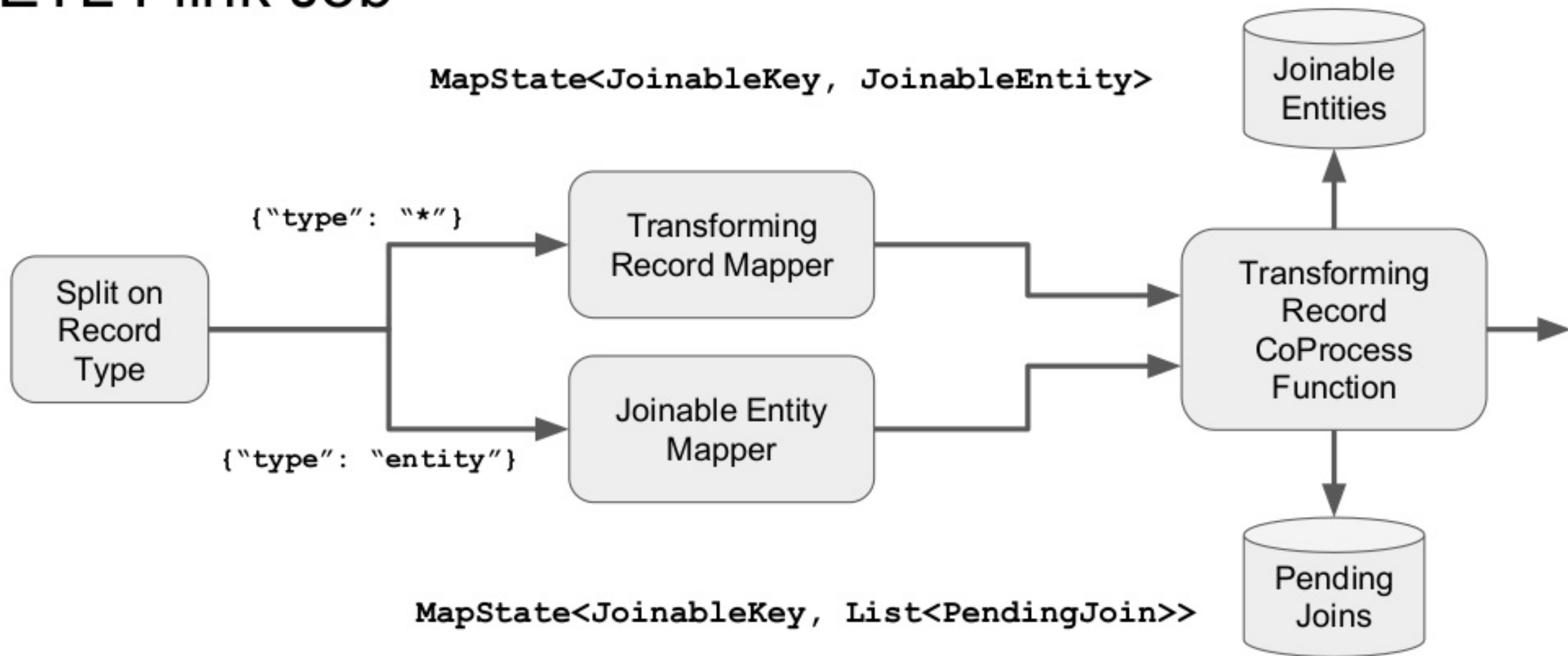


- Filter on Attributes
- Unwind Values

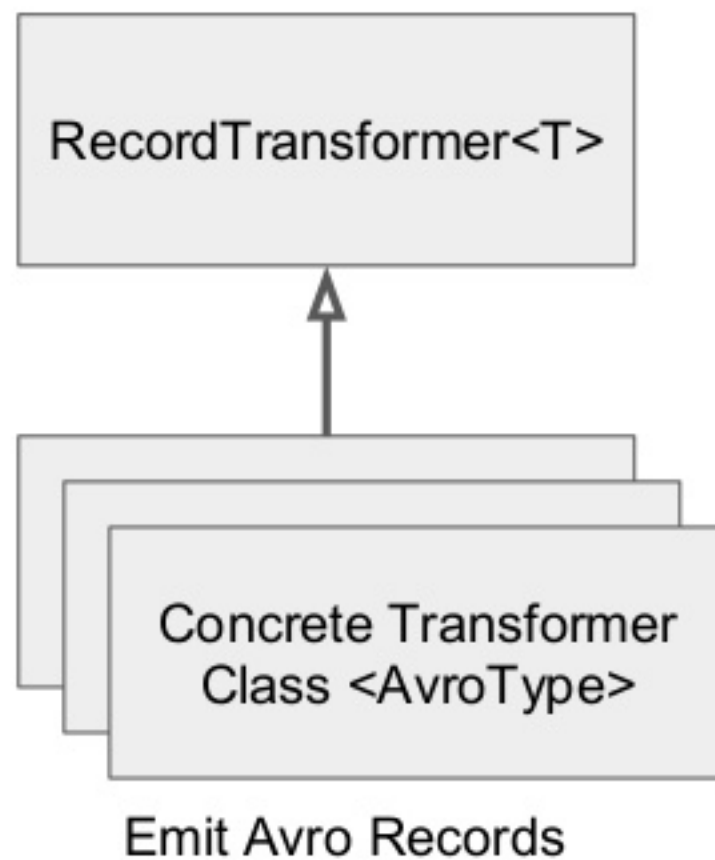
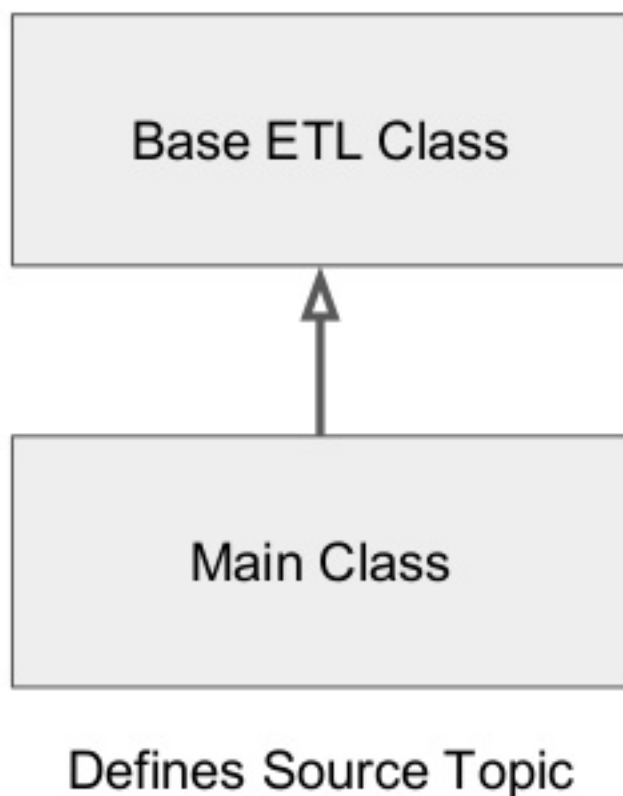
- Join Entity Data

- Translate Attributes
- Build Avro Object

ETL Flink Job



ETL Job - Framework



ETL Job - Main

```
@SpringBootApplication
```

```
public class LMSIngestEtlProgram extends IngestEtlProgram {  
  
    public static void main(String[] args) throws Exception {  
        IngestEtlProgram.etlFlinkMain(LMSIngestEtlProgram.class,  
                                       EtlSourceType.LMS, args);  
    }  
}
```


ETL Job - Record Transformer

`@Component`

1

```
public class QuizSubmissionTransformer extends RecordTransformer<QuizSubmission> {
```

```
    public QuizSubmissionTransformer(EntityJoinRegistry joinRegistry) {
        super(joinRegistry, QuizSubmission.class);
    }
```

`@Override`

3

```
public void configure(Builder<QuizSubmission> context) { /* */ }
```

`@Override`

2

```
public void flatMap(SourceRecordElement value, Collector<TransformingRecord> out) throws Exception { /* */ }
```

4

```
private static class Mapper implements MapFunction<SourceRecordElement, QuizSubmission> {
```

`@Override`

```
    public QuizSubmission map(SourceRecordElement value) throws Exception { /* */ }
```

```
}
```

```
}
```

ETL Job - Transformer - FlatMap

`@Override`

```
public void flatMap(SourceRecordElement value, Collector<TransformingRecord> out)  
throws Exception {
```

```
    if ("http://purl.imsglobal.org/caliper/v1/OutcomeEvent".equals(  
        stringProp(value.getRecord(), "event.@type"))) {
```

```
        out.collect(new TransformingRecord(getTransformContext(), value));
```

```
    }
```

```
}
```

ETL Job - Transformer - Configure

```
@Override
public void configure(Builder<QuizSubmission> context) {

    context.join("http://purl.imsglobal.org/caliper/v1/AssignableDigitalResource")
        .toField("activity")
        .on("entity.@id").eq("event.object.assignable.@id");

    context.join("http://purl.imsglobal.org/caliper/v1/lis/Membership")
        .toField("enrollment")
        .on("entity.member.@id").eq("event.generated.actor.@id")
        .and("entity.organization.@id").eq("activity.extensions.courseSection.subOrganizationOf.@id");

    context.map(new Mapper());
}
```

ETL Job - Transformer - Map Function

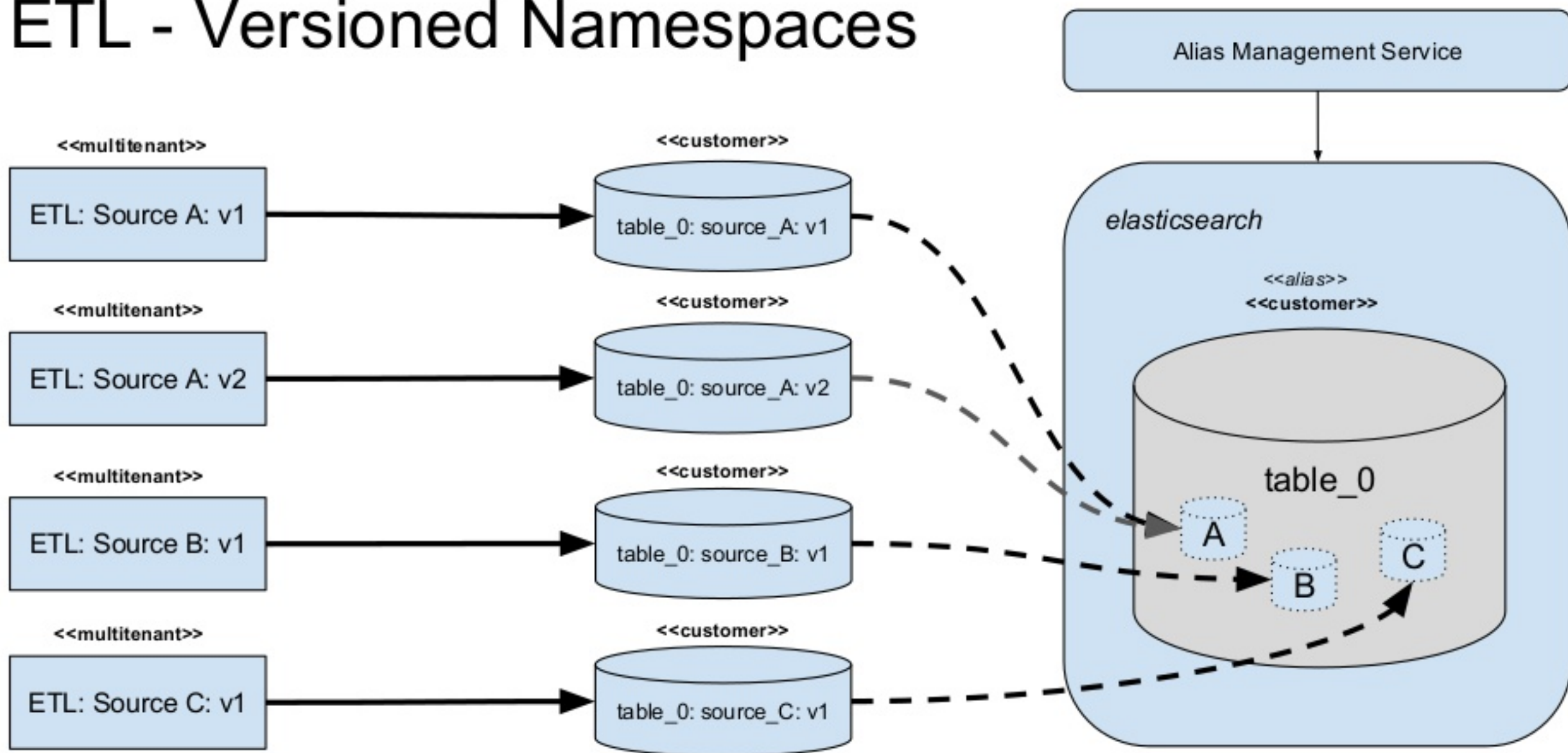
`@Override`

```
public QuizSubmission map(SourceRecordElement value) throws Exception {
    QuizSubmission.Builder val = QuizSubmission.newBuilder();
    ObjectNode rec = value.getRecord();

    val.setAction(stringProp(rec, "event.action"));
    val.setActivityId(stringProp(rec, "event.object.assignable.@id"));
    val.setActivityType(stringProp(rec, "activity.extensions.moduleType"));
    val.setAttemptId(stringProp(rec, "event.object.@id"));
    val.setCourseId(stringProp(rec, "activity.extensions.courseSection.subOrganizationOf.@id"));
    val.setDateAndTime(dateTimeProp(rec, "event.eventTime"));
    val.setRole(stringProp(rec, "enrollment.roles"));
    val.setScore(doubleProp(rec, "event.generated.totalScore", 0d));
    val.setStudentId(stringProp(rec, "event.generated.actor.@id"));

    return val.build();
}
```


ETL - Versioned Namespaces



Flink Bootstrapping Source

Range:

`{seedStart, seedEnd}`



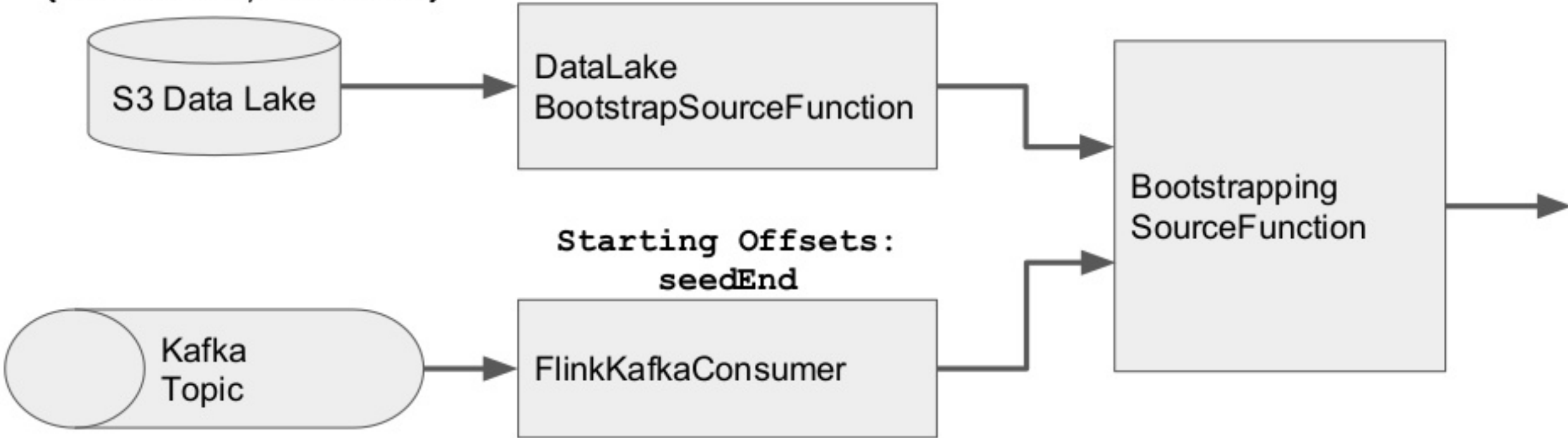
DataLake
BootstrapSourceFunction

Starting Offsets:
`seedEnd`

Bootstrapping
SourceFunction



FlinkKafkaConsumer



Flink Bootstrapping Source

`@Override`

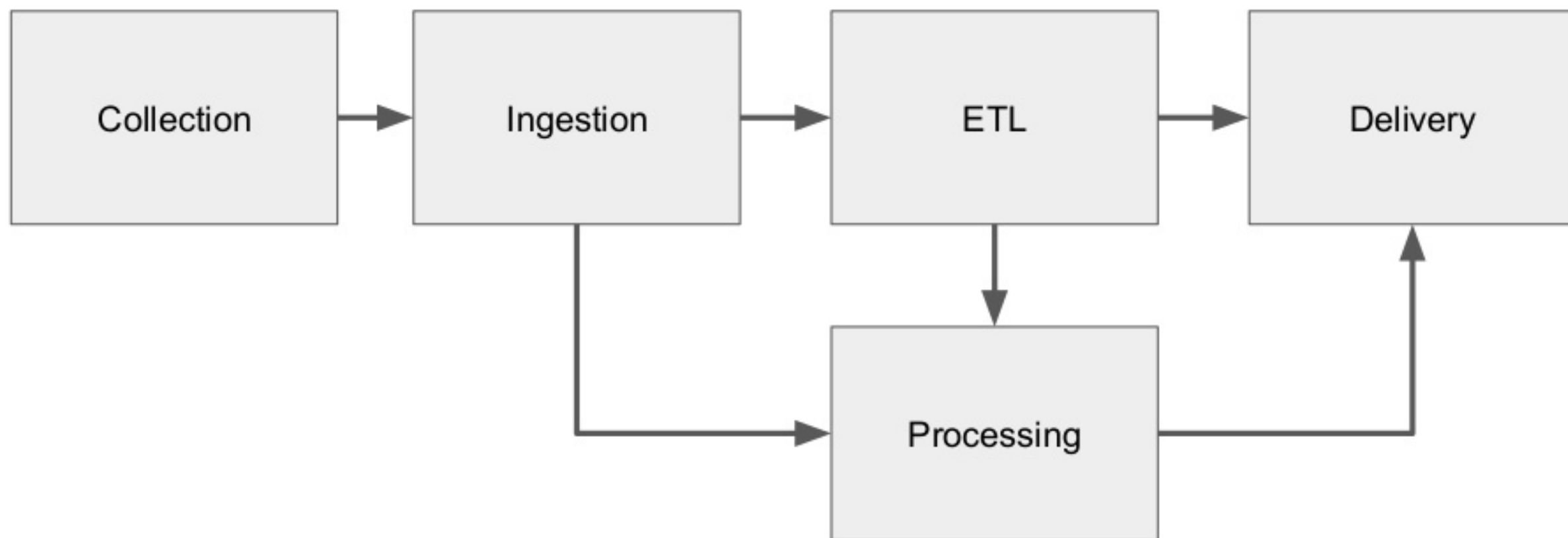
```
public void run(SourceContext<T> ctx)
    throws Exception {

    while (running) {
        if (seeding.get()) {
            if (bootstrapSource != null) {
                bootstrapSource.run(ctx);
            }

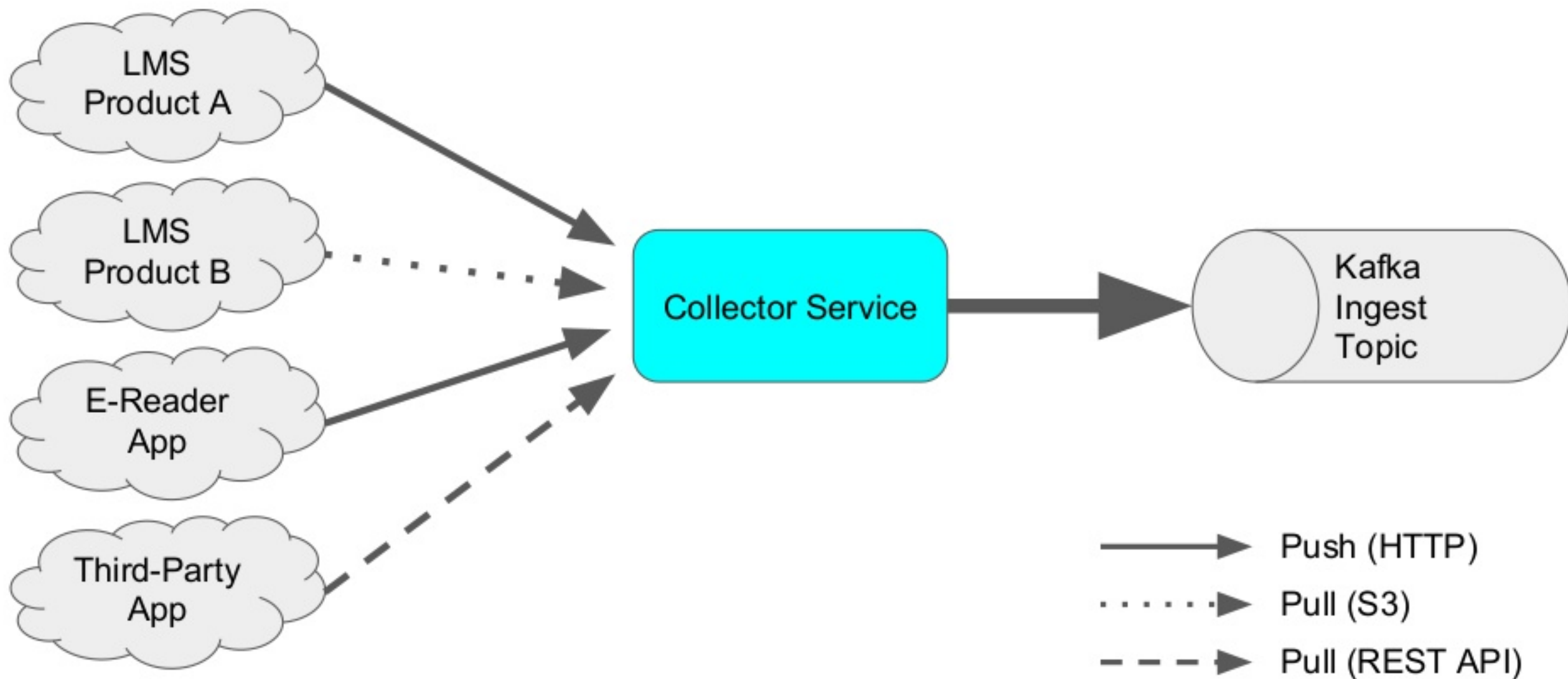
            seeding.set(false);

            // yield momentarily to allow other parallel seeding tasks time to complete
            Thread.sleep(5_000);
        } else {
            streamSource.run(ctx);
        }
    }
}
```

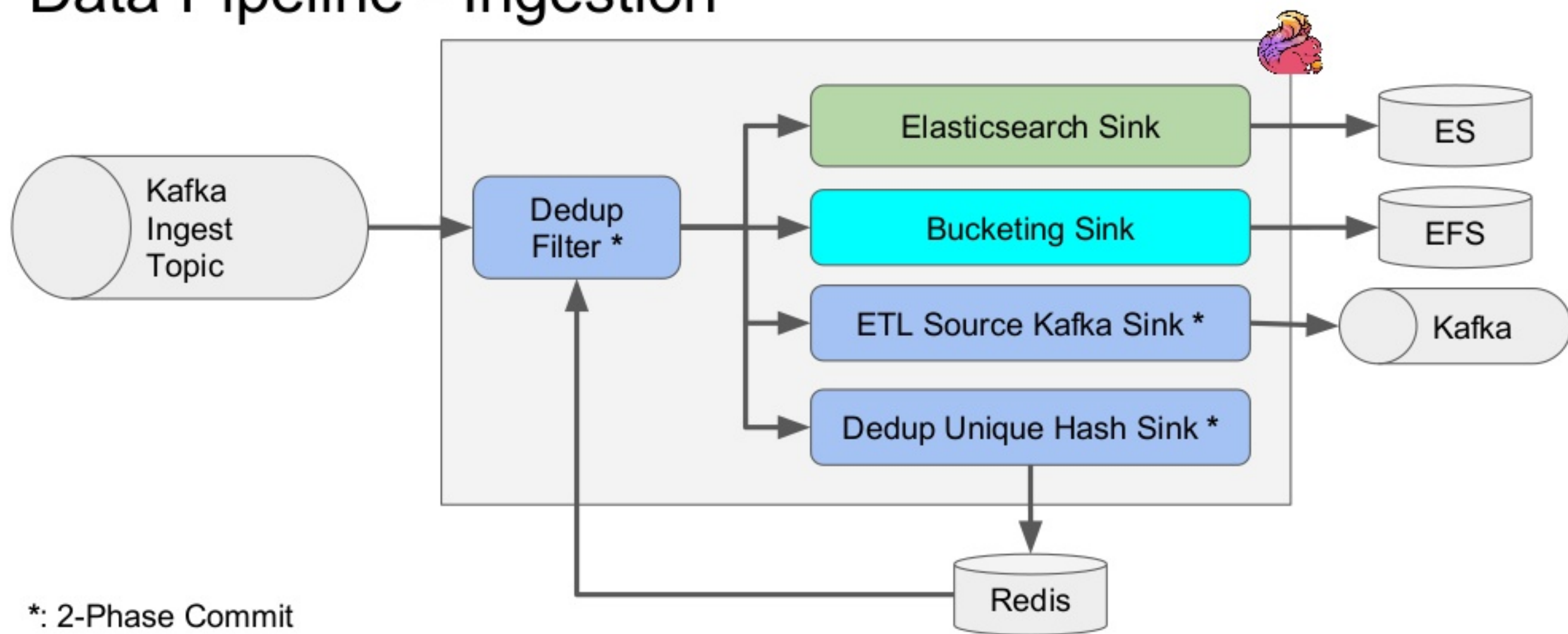
Data Pipeline - Overview



Data Pipeline - Collection

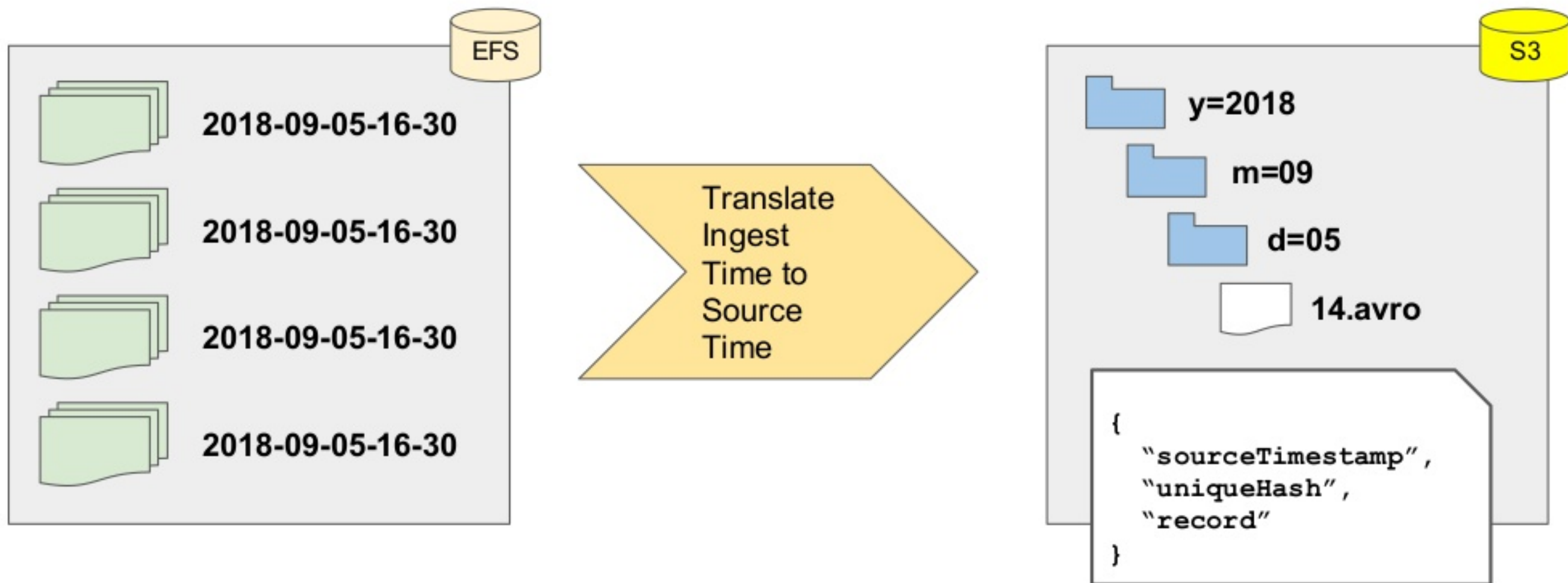


Data Pipeline - Ingestion

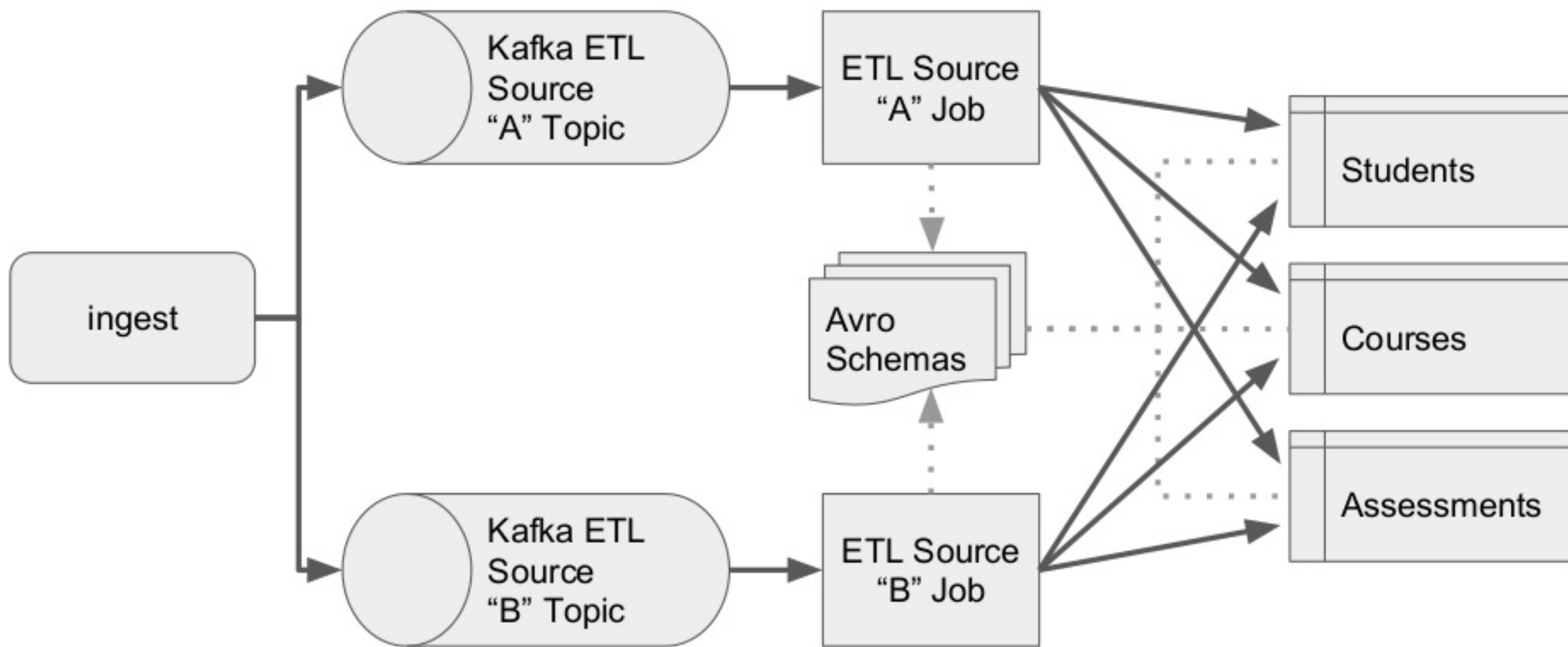


*: 2-Phase Commit

Data Pipeline - Ingestion (Raw Data Lake)



Data Pipeline - ETL



Data Pipeline - “Processing”

- Sessionization Jobs: “Time on Task”, “Concurrent Active Sessions” Reports
- Daily / YTD Aggregations: Student, Course, School, District
- Future: CEP - Detect Cheating, etc

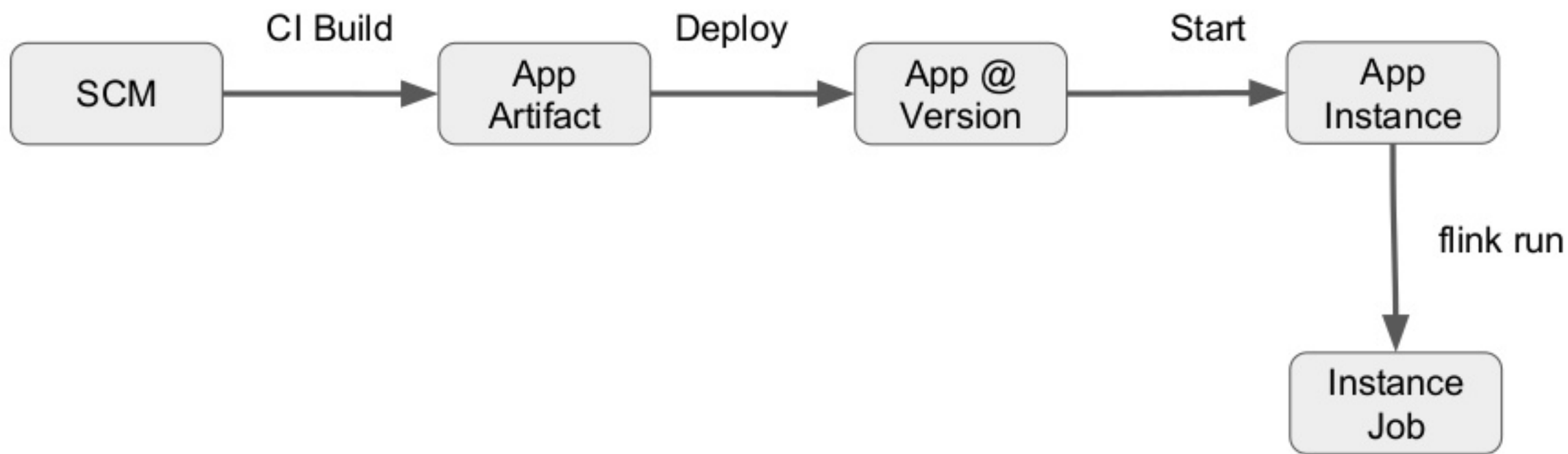
Data Pipeline - Delivery

- Real-time reports created / delivered via Elasticsearch indices
- BI Analytics / exploration via SQL on S3 data in parquet buckets by timestamp
 - Presto / Athena
 - Drill

Flink Deployment Model

- **DataStream App**
 - CI builds versioned artifacts
 - Deployable to environments via descriptor
- **DataStream App Instance (1..*)**
 - Produces output to namespaced sinks
 - Actions:
 - **Suspend**: cancel with savepoint
 - **Resume**: start from savepoint
 - **Recover**: start from last external checkpoint
 - **Upgrade**: update App version to currently deployed artifact
- **DataStream App Instance Job (1..*)**
 - Link between App Instance and Flink Job Instance


Flink Deployment Model



DS App Deployment Descriptor



```
{  
  "type": "datastream-app",  
  "description": "ETL job for edgenuity-lms data sources",  
  "multitenant": true,  
  "versioned": true,  
  "seedingEnabled": true,  
  "seedingType": "datasource_type",  
  "enabledSeedingTargets": [  
    "edgenuity_lms"  
  ]  
}
```

Deploying


Vulcan
[Deploys](#)
[Builds](#)
[Stacks](#)
[Blockers](#)
[jared.stehler](#)

New Deploy

Artifact Type	datastream-app
Artifact	ingest-etl-edgenuity-lms-job - [datastream-app]
Stack	master-prod1
Existing Deploy	1.0.26 - 2018-08-16 12:05 pm - hans.parra
Version	1.0.34

 Descriptor
  Diff

[+ Advanced Options](#)

Deploy


DS App Management UI

crusher-prime-job (1.2.76)
Start

ingests raw data from collector into data lake

unversioned (1.2.76)

suspend
delete

97,926


RUNNING

global-ingest-v3

Last updated at 2018-08-27 10:13 pm by system

[info](#) [history](#)

ingest-etl-edgenuity-lms-job (1.0.26)
Start

ETL job for edgenuity-lms data sources

20180816160759 (1.0.26)

recover
delete

CANCELED

ingest-etl-edgenuity-lms-job

Last updated at 2018-08-17 12:19 pm by jared.stehler

[info](#) [history](#)

DS App - Start New Instance

Master ProfilesTools

Start App Instance

Start Fresh

Start this app from the beginning of its input source, optionally with seeding data from the data lake.

Job Parallelism

1

Seeding Target

edgenuity_lms

Seeding Date Range

05/31/2014 20:00 - 08/27/2018 23:10

Included Envs

☐ deepthought:deepthought1 (51001588)
☐ deepthought:deepthought2 (256558876)
☒ deepthought:deepthought3 (70631567)

Estimated Record Count

378,192,031

- deepthought (378192031)
 - deepthought1 (51001588)
 - deepthought2 (256558876)
 - deepthought3 (70631567)

Start with Seeding

DS App Management UI

crusher-prime-job (1.2.76)

Start

unversioned (1.2.76)

global-ingest-v3

suspend

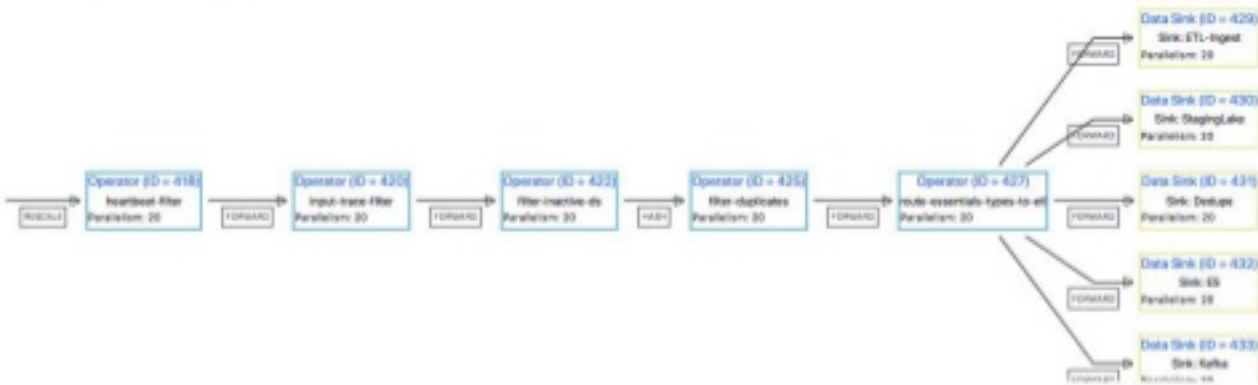
delete

97,926

RUNNING

Last updated at 2018-08-27 10:13 pm by system

[info](#)
[history](#)



Sources

KAFKA

ingest-prime

Sinks

DS App Management UI

crusher-prime-job (1.2.76)

Ingests raw data from collector into data lake

Start

unversioned (1.2.76)

global-ingest-v3


Last updated at 2018-08-27 10:13 pm by system

[info](#)
[history](#)

suspend

delete

97,926



RUNNING

RUNNING (1.2.76) (2018-08-27 10:13 pm -)

20 2018-08-27 10:08 pm

CANCELED (1.2.76) (2018-08-25 07:09 am - 2018-08-25 07:09 am)

20 2018-08-25 12:19 am

DS App - Flink Job Config Params - Versioned

User configuration	
accountId	system
dataVersion	20180827174235
startedBy	hans.parra
sourceType	EDGENUITY_LMS
program	ingest-etl-edgenuity-lms-job
envId	system
jobGroup	ingest-etl
version	1.0.34
uniqueId	ingest-etl-edgenuity-lms-job-system-system--20180827174235

DS App - Flink Job Config Params - Singleton

User configuration	
accountId	system
dataVersion	unversioned
startedBy	jared.stehler
program	crusher-prime-job
envId	system
version	1.2.75
uniqueId	crusher-prime-job-system-system--unversioned

Flink Pipeline Monitoring

- `rate(write_records{unique_id="crusher-prime-job-system-system--unversioned", vertex_type="source"} [5m])`
- `max(flink_jobmanager_job_lastcheckpointduration{jobname=~"global-ingest-.*"}) / 1000.0`
- `sum(rate(flink_jobmanager_job_fullrestarts[10m])) BY (jobname) > 0`

Flink Pipeline Monitoring



Thanks!



Jared Stehler | @jstehler