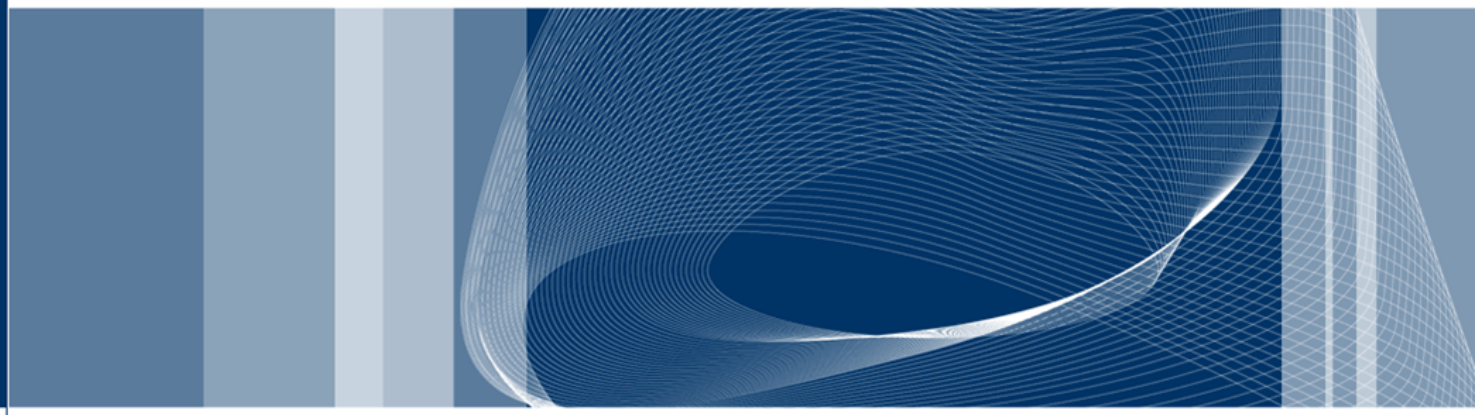
POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica

Anno Accademico 2013 - 2014



Avoiding CRUD operations lock-in in NoSQL databases: extension of the CPIM library

Candidato: Fabio Arcidiacono (799001)

Relatore: Prof.ssa Elisabetta Di Nitto

Correlatore: Ing. Marco Scavuzzo

Data management systems

RDBMS

- Well structured data
- Relational model
- ACID transactions
- Vertical scaling
- **SQL**

NoSQL

- Non-structured data
- Various data models
- BASE properties
- Horizontal scaling
- **Proprietary API**

NoSQL Common language approaches

Meta-model

- Apache MetaModel
- SOS platform

SQLification

- Apache Phoenix
- UnQL
- Native support

ORM

- Kundera
- PlayORM
- Spring-data
- Apache Gora

Work objectives

Contribute to the open source project Kundera

Integrate Kundera in the CPIM library

Integrate the migration and synchronization system Hegira

Evaluation

Work objectives

Contribute to the open source project Kundera

Integrate Kundera in the CPIM library

Integrate the migration and synchronization system Hegira

Evaluation

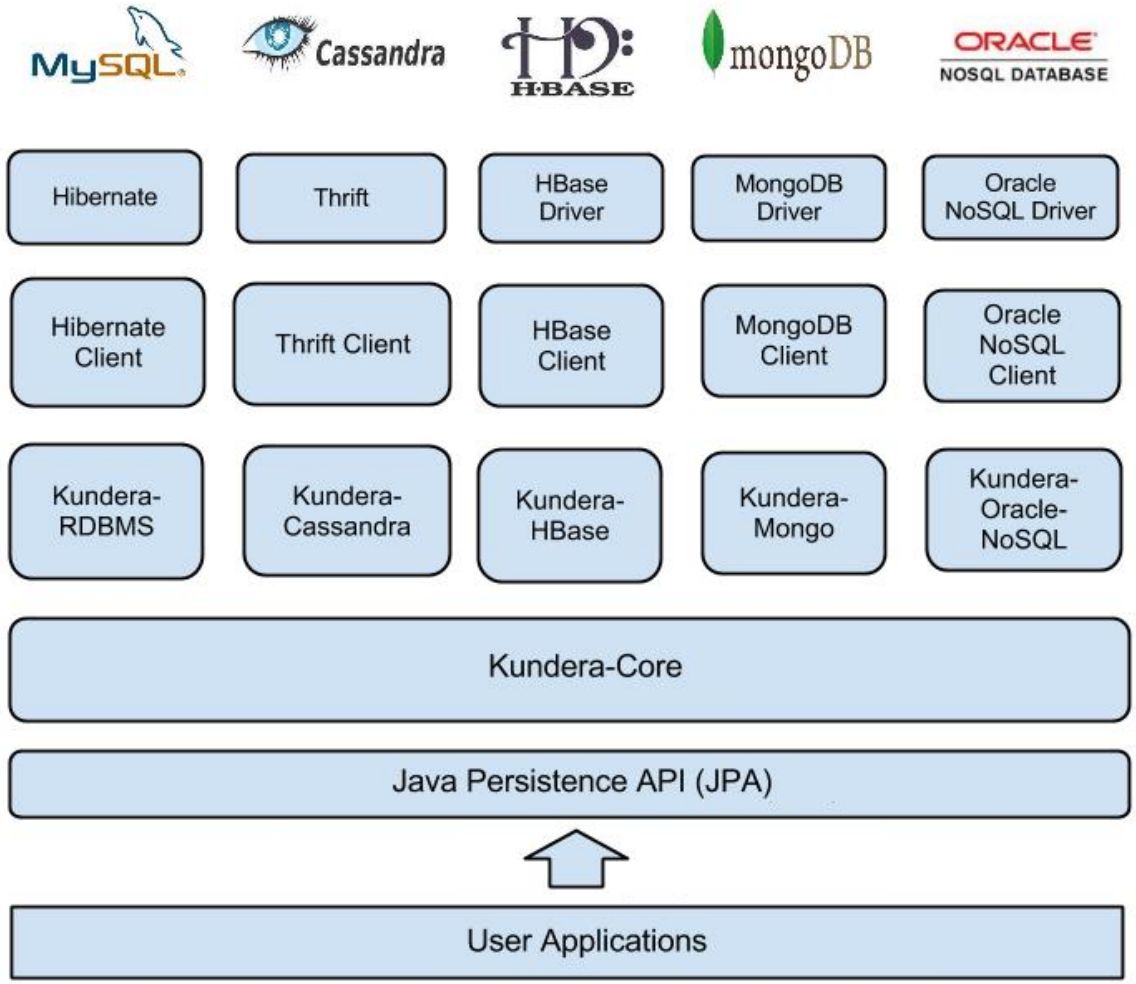
Kundera

A JPA 2.1 ORM Library for NoSQL databases

- ORM operation (through *EntityManager* interface)
- JPQL queries (DELETE and UPDATE)

On-premises databases:

- Cassandra
- HBase
- MongoDB
- Oracle NoSQL
- Redis
- Neo4j
- Couchdb
- Elastic Search
- MySQL



Why Kundera

- Open source
- Developed with extensibility as primary goal
- Support to many different NoSQL databases
- Polyglot persistency
- In the field since 2010 with an active community
- Already used in production

Contributions to Kundera

Open Two newly developed clients

- Azure Tables[1]
- GAE Datastore[2]

Paradigm shift

- Off-premises databases → DaaS solutions
- Merged Bug fix Kundera deploy on PaaS

1: https://github.com/deib-polimi/kundera-azure-table

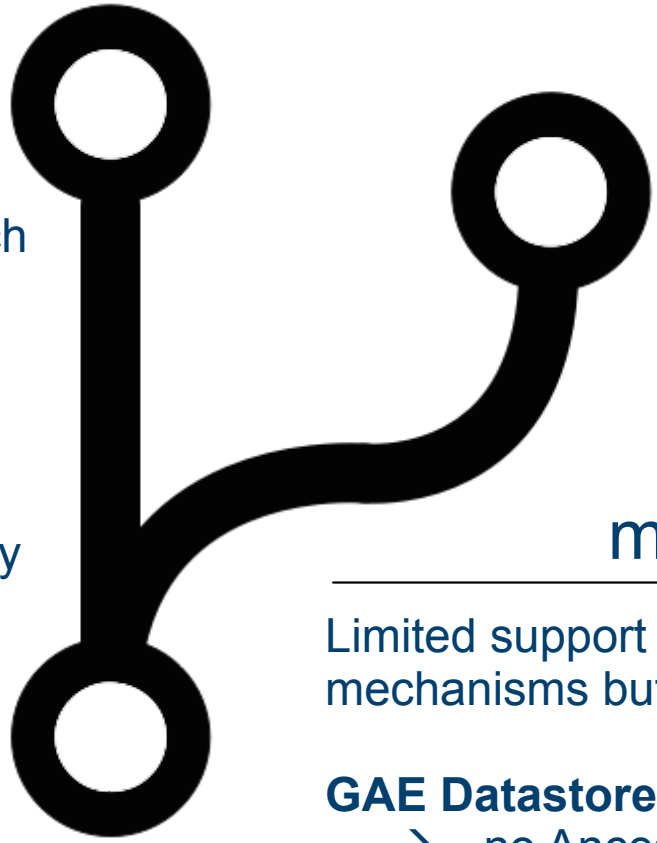
2: https://github.com/deib-polimi/kundera-gae-datastore

Developed clients

master

Exploit consistency mechanisms as much as possible

GAE Datastore
→ no Ancestor Path support
Azure Tables
→ manage partition key and row key



migration

Limited support to consistency mechanisms but achieve interoperability

GAE Datastore
→ no Ancestor Path support
Azure Tables
→ fix partition key to table name

Work objectives

Contribute to the open source project Kundera

Integrate Kundera in the CPIM library

Integrate the migration and synchronization system Hegira

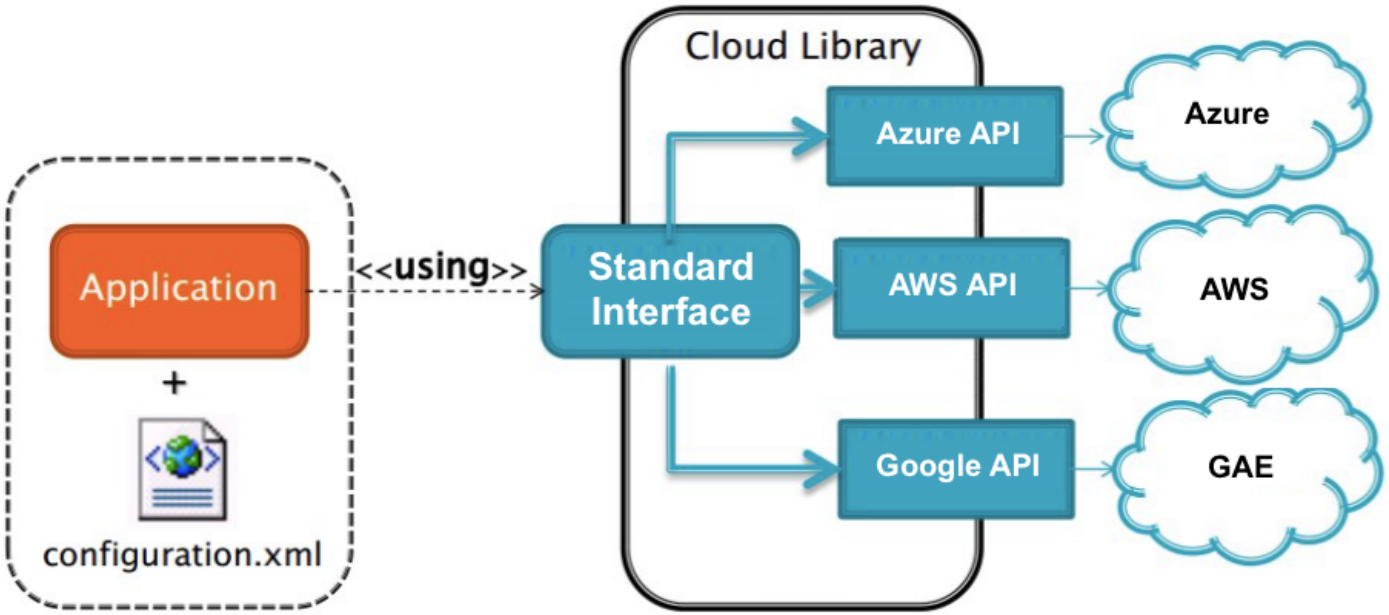
Evaluation

CPIM

Abstract application logic from the specific PaaS Provider to overcome the vendor lock-in

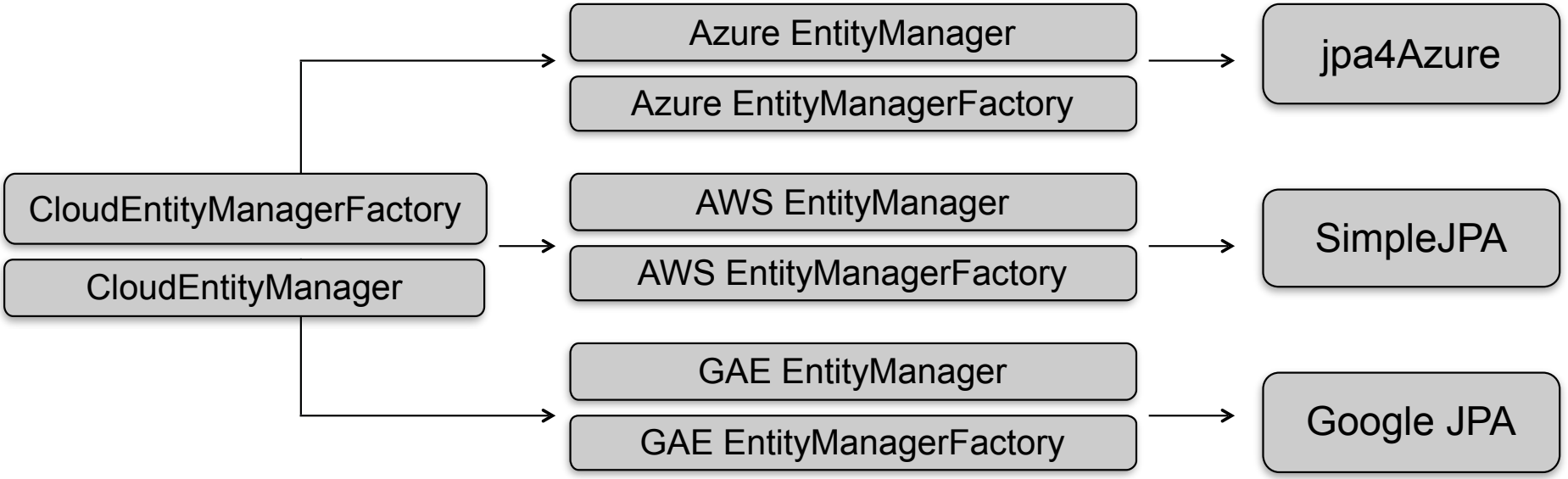
Many supported services:

- Blob
- NoSQL
- Memcache
- Queue
- Mail
- SQL



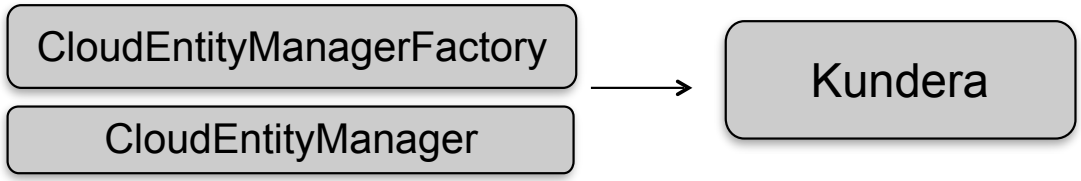
Original CPIM NoSQL service implementation

- Many JPA providers
- Duplicated code
- No complete code portability
- Choice of the NoSQL database strictly bounded to the cloud provider (e.g. App Engine → Datastore)
- Limited NoSQL databases support



Kundera integration

- Single persistence provider
- Complete code portability
- NoSQL support inherited by Kundera
- Easier Configuration through standard persistence.xml



Work objectives

Contribute to the open source project Kundera

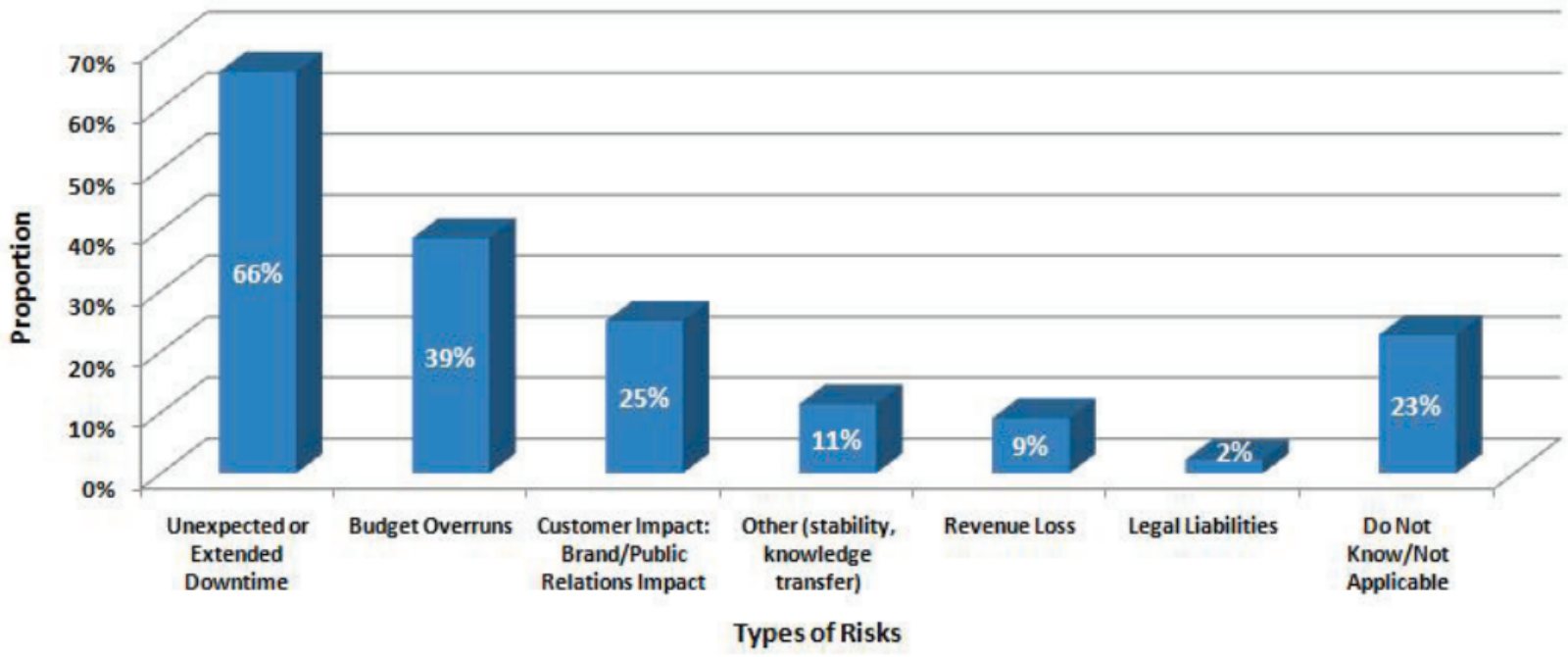
Integrate Kundera in the CPIM library

Integrate the migration and synchronization system Hegira

Evaluation

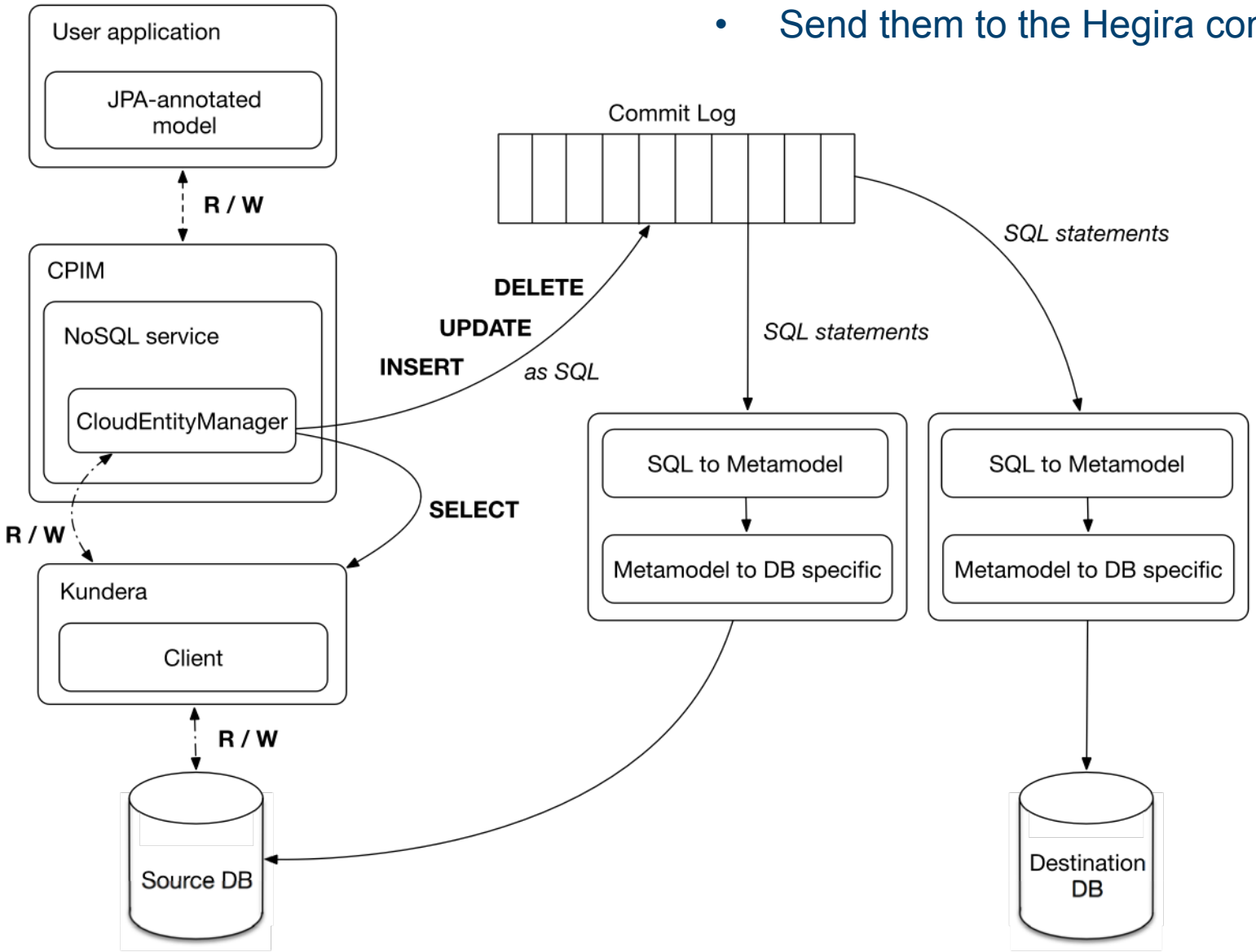
Data migration

- move application to another cloud provider
- move data to a database that better fit requirements
- load balancing, system expansion, failure recovery, costs, etc.
- modern computer systems are expected to be up continuously
- data synchronization between the two involved systems



Hegira support

- Intercept transparently user operations (DMQ)
- Translate operations to SQL statements
- Send them to the Hegira commit-log



Work objectives

Integrate Kundera in the CPIM library

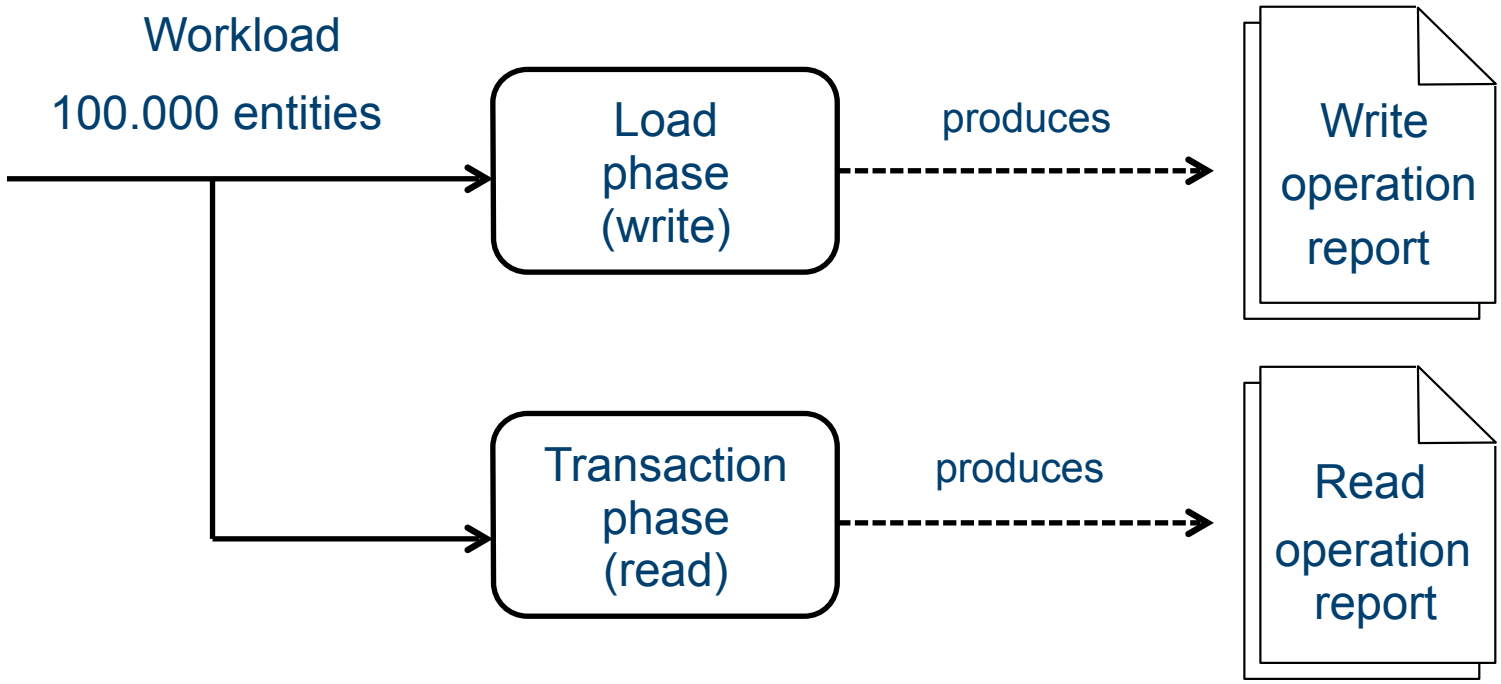
Contribute to the open source project Kundera

Integrate the migration and synchronization system Hegira

Evaluation

YAHOO! Cloud Serving Benchmark

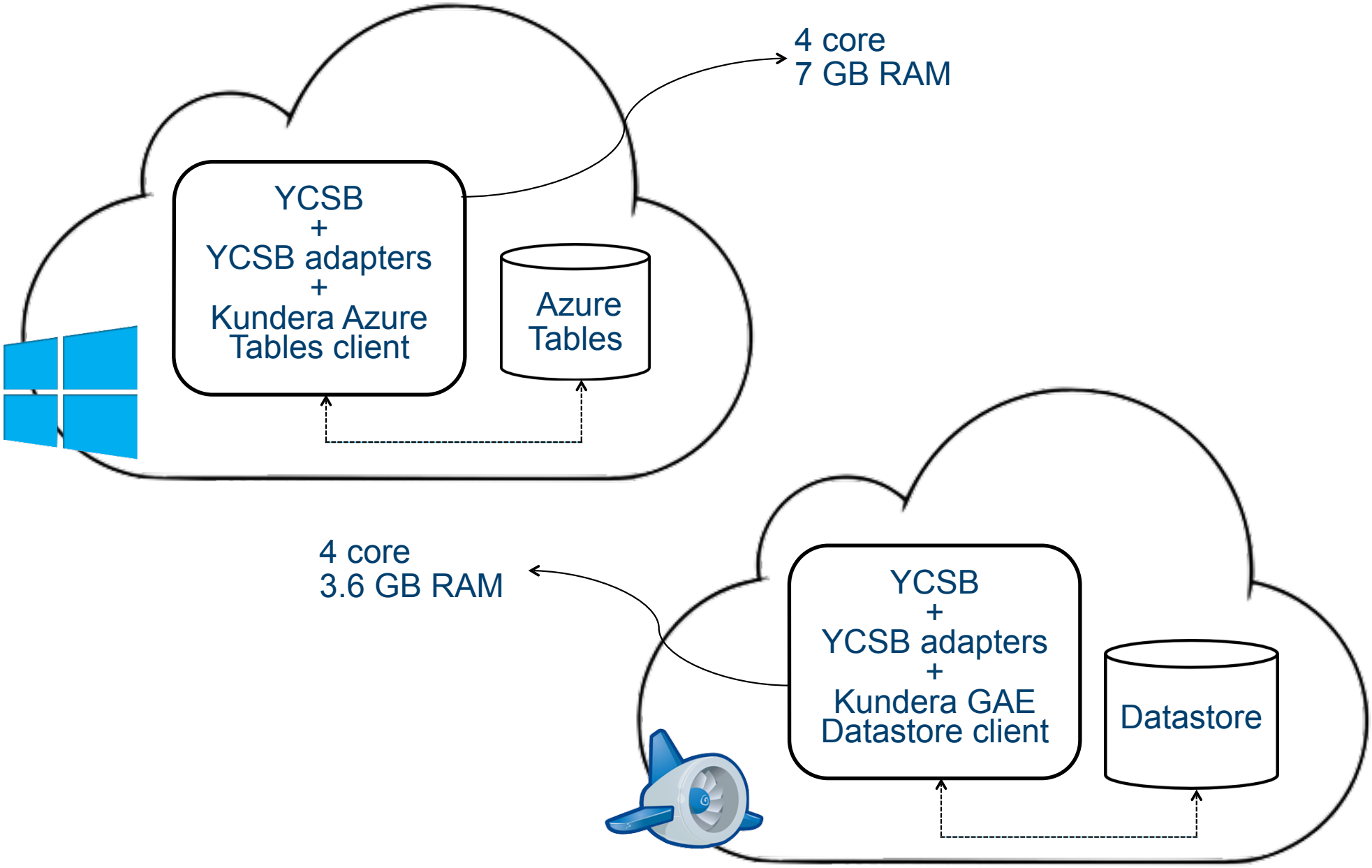
Framework for evaluating the performance of different NoSQL databases



Compare Kundera client w.r.t. the use of low-level API for the same operations

- Development of new adapter for operations through Kundera
- Development of new adapter for operations through the low-level API

Environment setup



Results comparison

Azure Tables

	Read latency	Read throughput	Write latency	Write throughput
Kundera	42,44 *ms*	689,67 *ops/sec*	40,701 *ms*	707,12 *ops/sec*
low-level API	36,74 *ms*	787,22 *ops/sec*	38,809 *ms*	758,54 *ops/sec*
overhead	13,43 *%*	12,39 *%*	4,75 *%*	6,78 *%*

Google Datastore

	Read latency	Read throughput	Write latency	Write throughput
Kundera	139,13 *ms*	212,74 *ops/sec*	151,159 *ms*	194,64 *ops/sec*
low-level API	132,36 *ms*	222,5 *ops/sec*	150,018 *ms*	198,67 *ops/sec*
overhead	4,36 *%*	4,39 *%*	0,76 *%*	2,03 *%*

Conclusions

Contributions:

- Integration of Kundera in CPIM library
- New Kundera clients to support Google Datastore and Azure Tables
- Hegira integration in the CPIM library

Future work:

- Compare developed client performance with the ones of the other client developed by Kundera team

THANK YOU