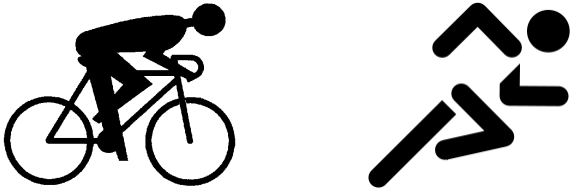
Offload, Transform, and Present - the New World of Data Integration

Michael Rainey

Introduction

- Michael Rainey - Technical Advisor
- Spreading the good word about gluent products with the world
  - Data Integration expertise
  - Oracle ACE Director
  - @mRainey

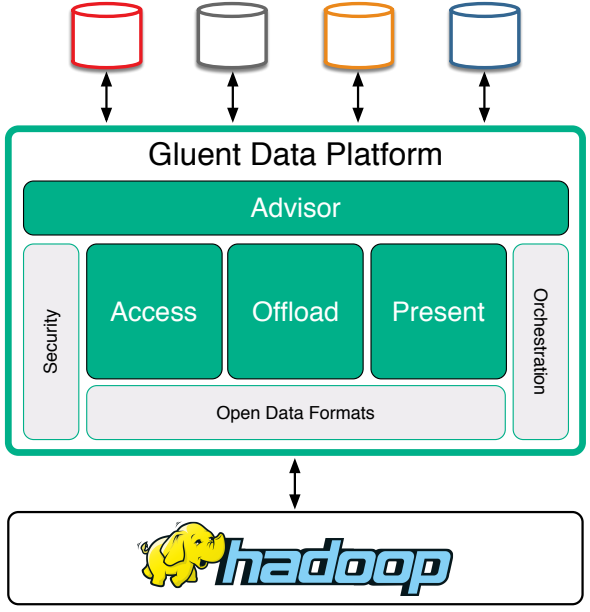


gluent.
we liberate enterprise data

About Gluent

Globally distributed team with a deep background in building high performance enterprise applications and systems …

Now on a mission to liberate enterprise data.



About Gluent

Globally distributed team with a deep background in building high performance enterprise applications and systems …

Now on a mission to liberate enterprise data.

Recognized by Gartner in their Cool Vendors in Data Management, 2017 report!

Download (no registration needed): https://gluent.com/cool-vendor-2017/

Listed as one of the "Top 10 Coolest Big Data Startups of 2017" by CRN

gluent.com/gluent-recognized-in-top-10-coolest-startups-of-2017-on-crn

2nd place in Strata+HadoopWorld Startup Showcase 2017.

(Demo video at gluent.com)

Gluent Webinars - JULY 2017

Gluent is running one-hour webinars each Wednesday in July beginning tomorrow, July 12. See details below and register for your free spot today!

Apache Impala Internals

Speaker: Tanel Poder, Gluent

Wednesday, July 19 @ 12 PM CDT

gluent.com/event/gluent-webinar-apache-impala-internals-with-tanel-poder

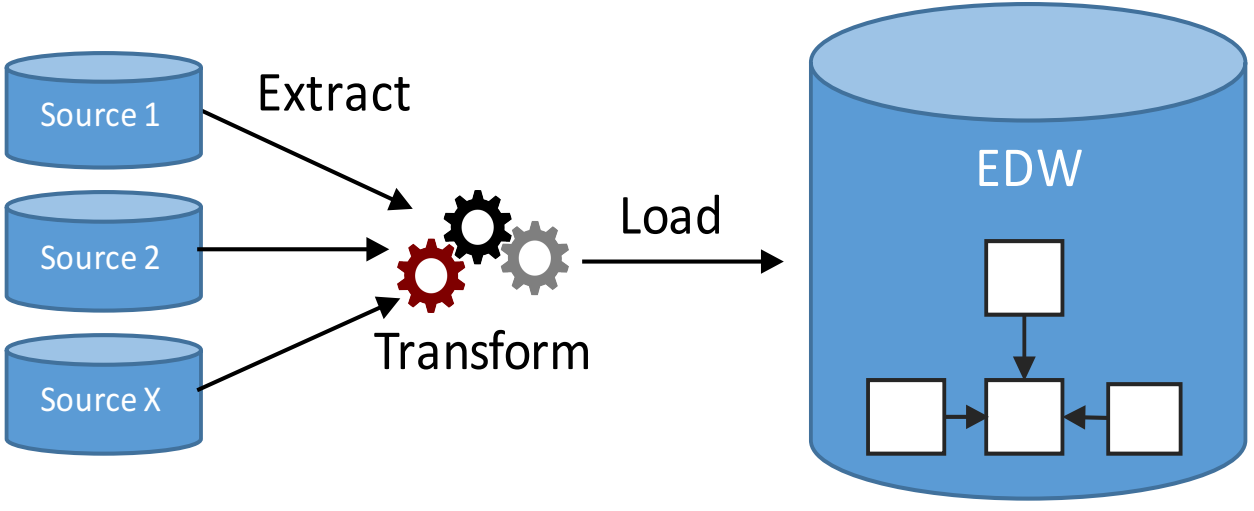
Building an Analytics Platform with Oracle & Hadoop

Speakers: Gerry Moore & Suresh Irukulapati, Vistra Energy

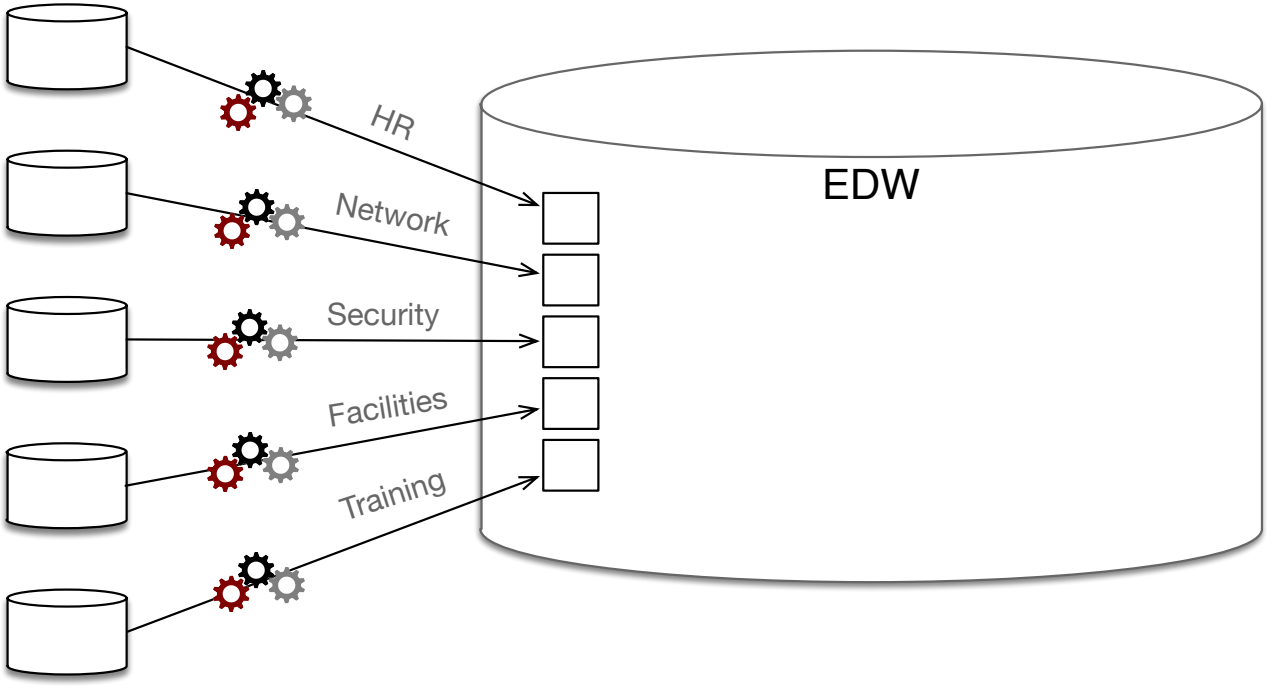
Wednesday, July 26 @ 9 AM CDT

gluent.com/event/gluent-webinar-building-an-integrated-analytics-platform-with-oracle-and-hadoop

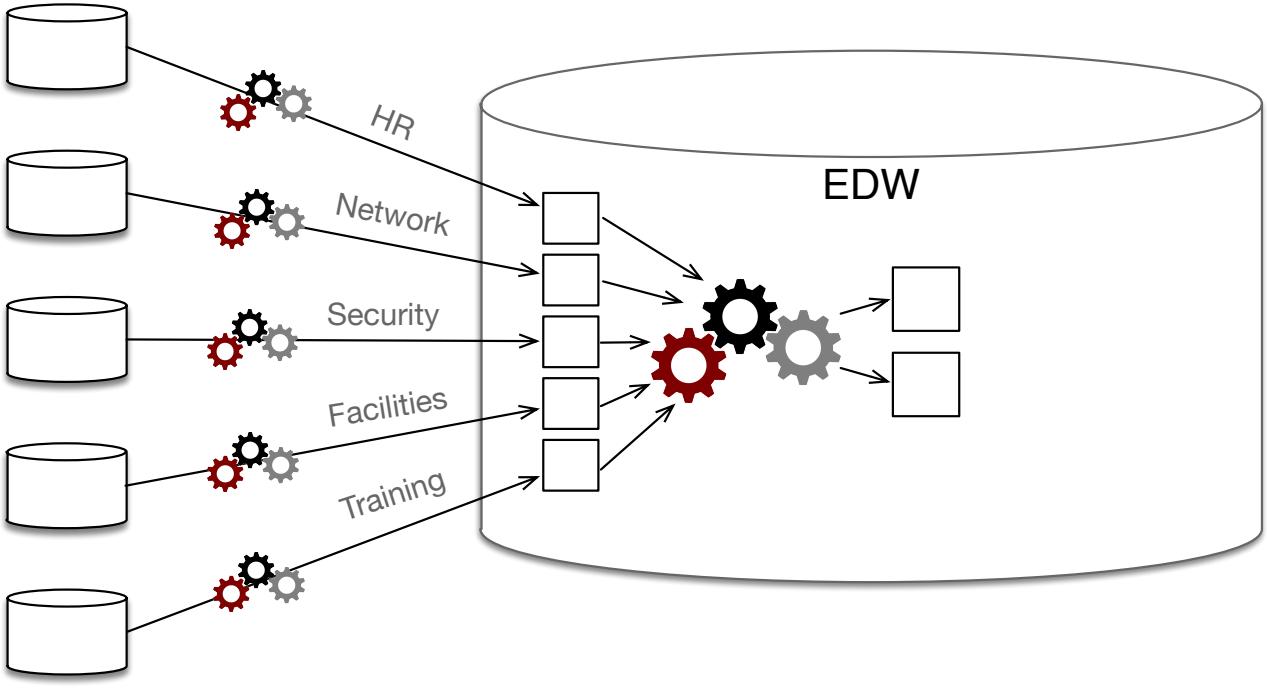
Extract Transform Load (ETL)



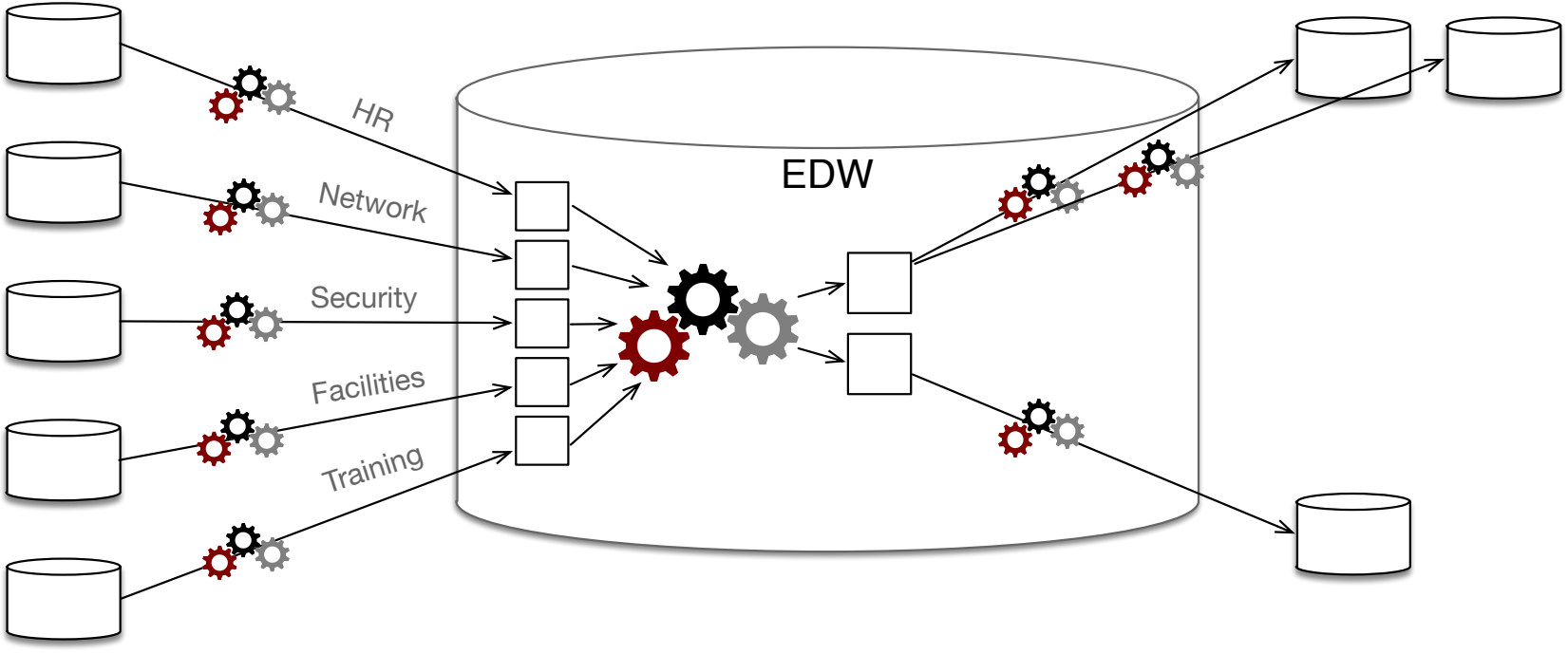
ETL Examples - employee data



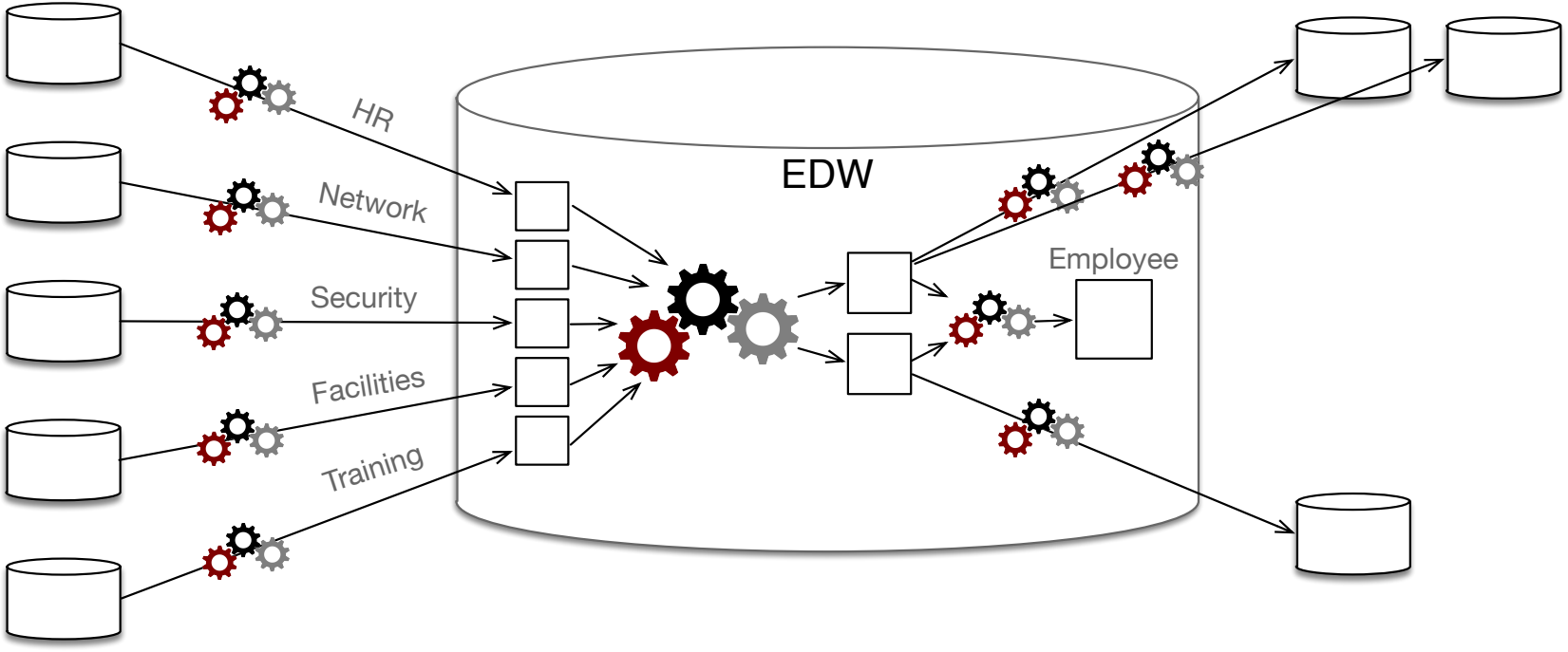
ETL Examples - employee data



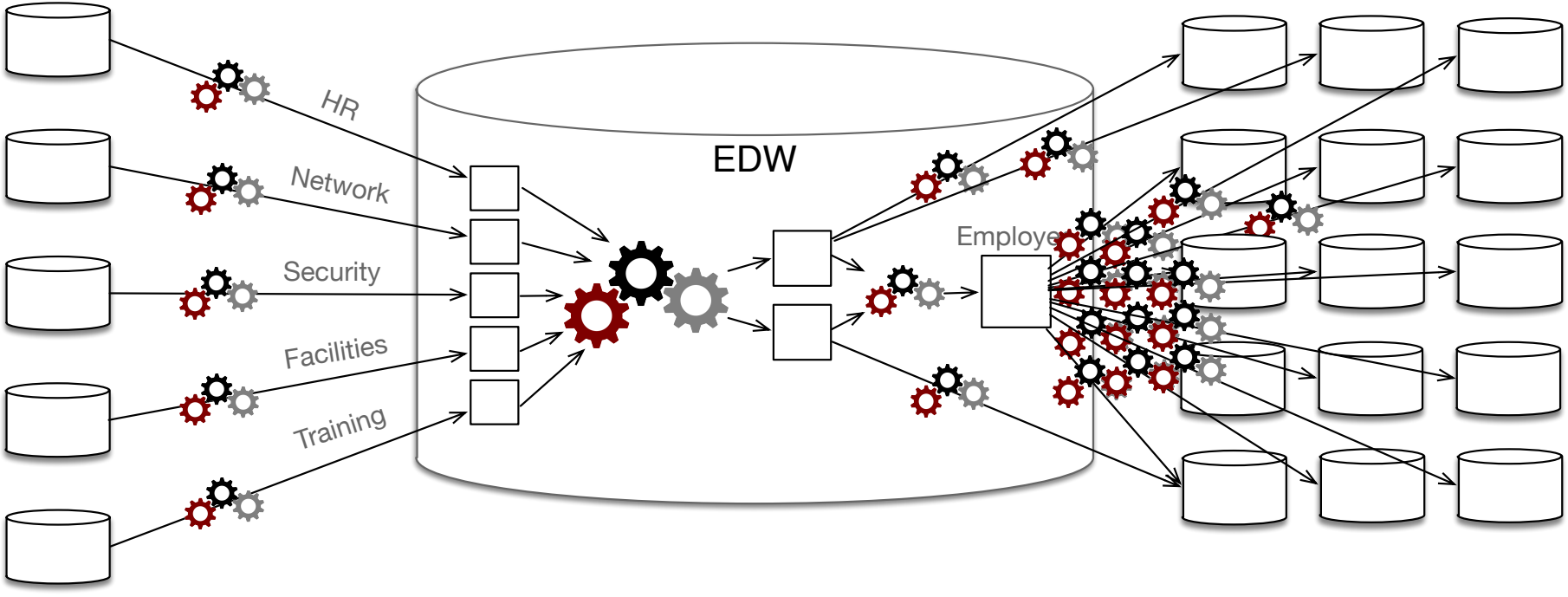
ETL Examples - employee data



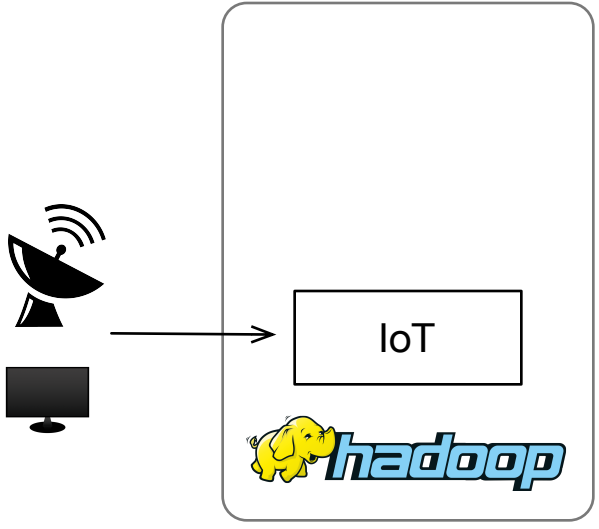
ETL Examples - employee data



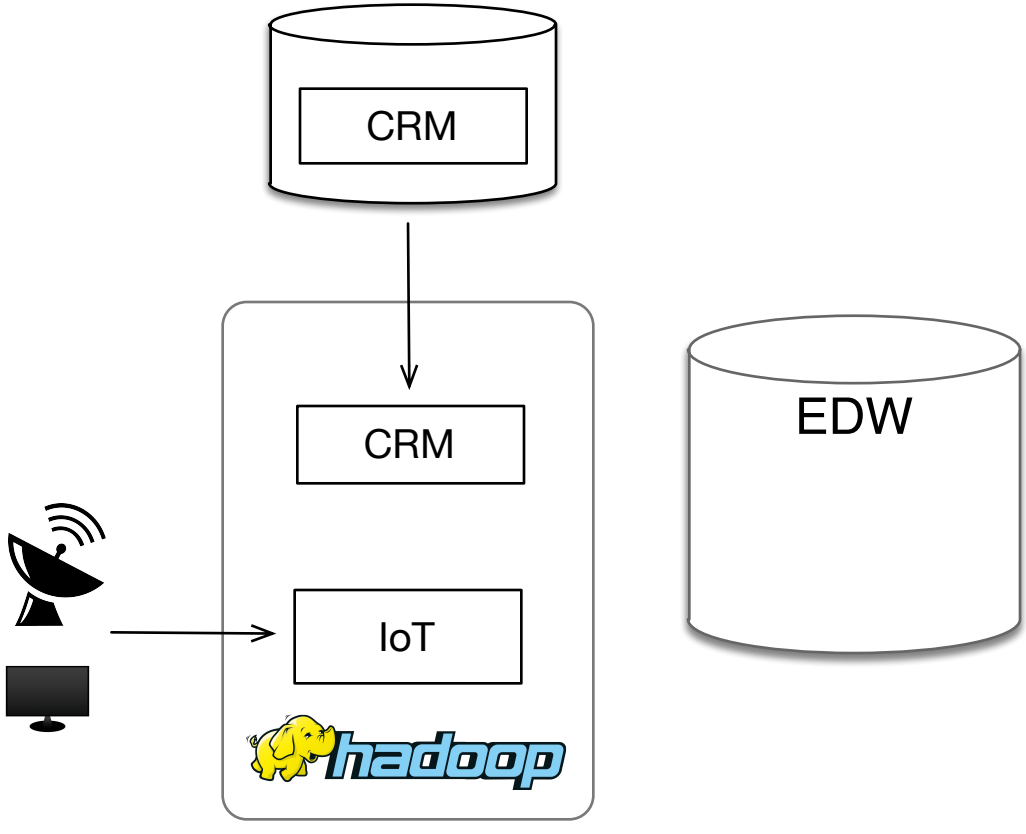
ETL Examples - employee data



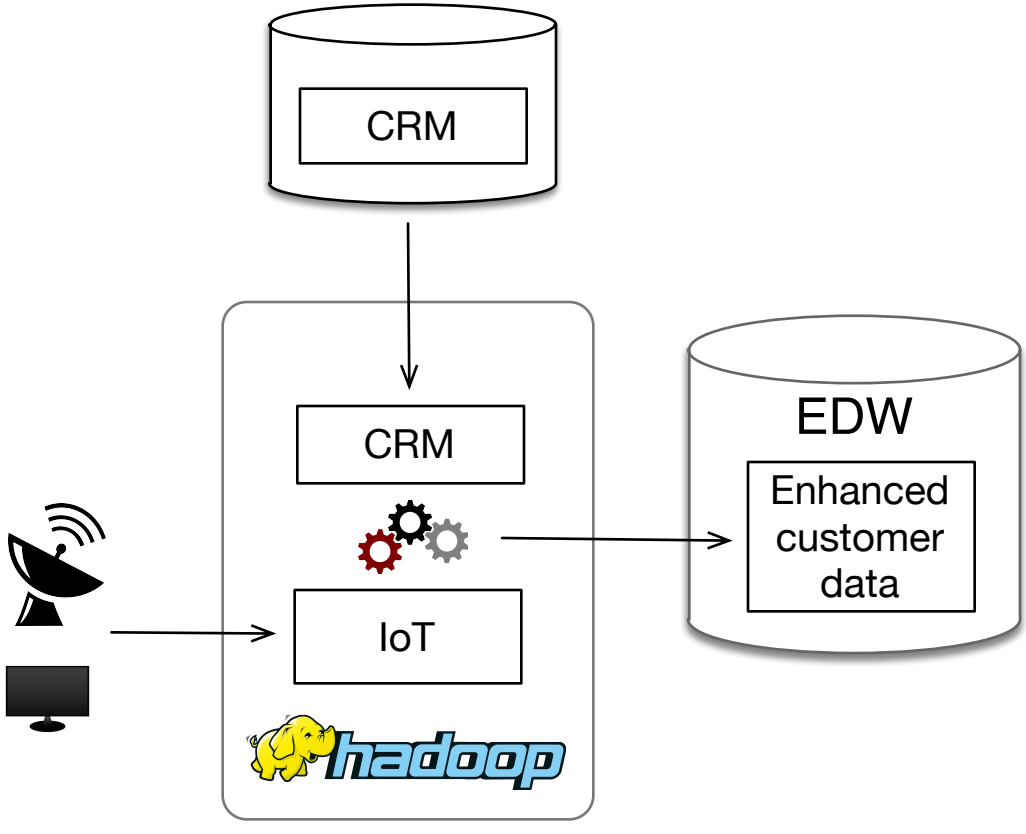
ETL Examples - IoT from set-top boxes



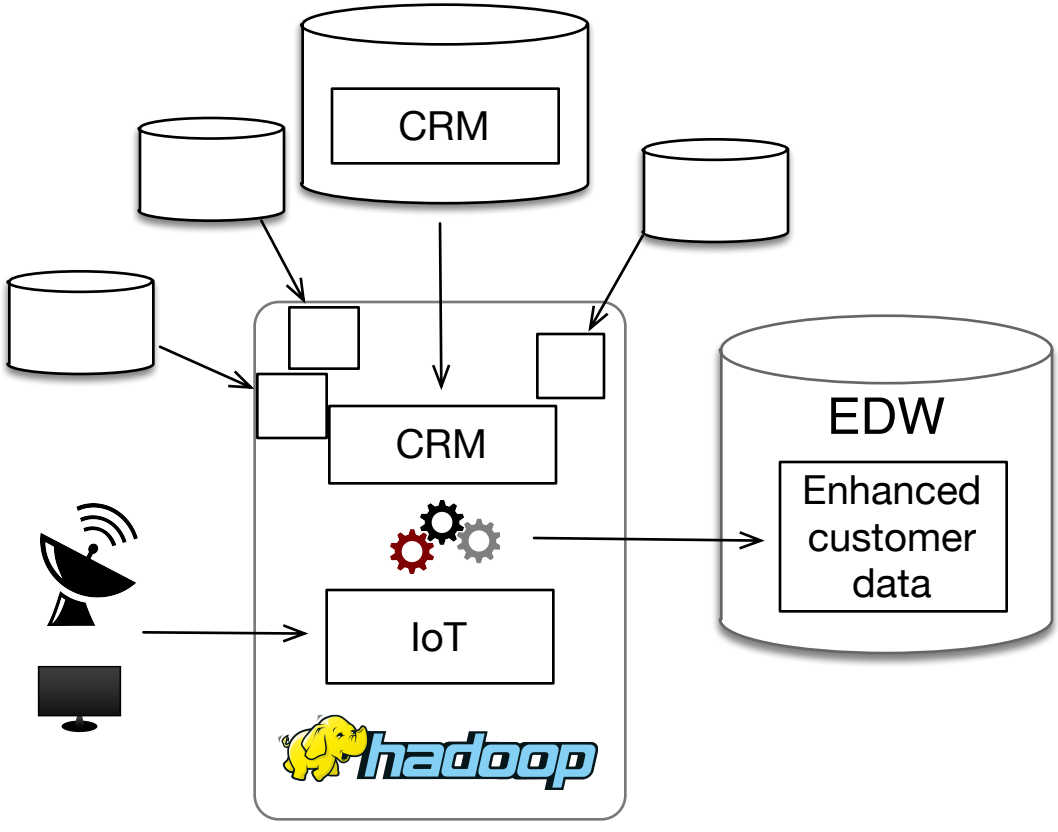
ETL Examples - IoT from set-top boxes



ETL Examples - IoT from set-top boxes

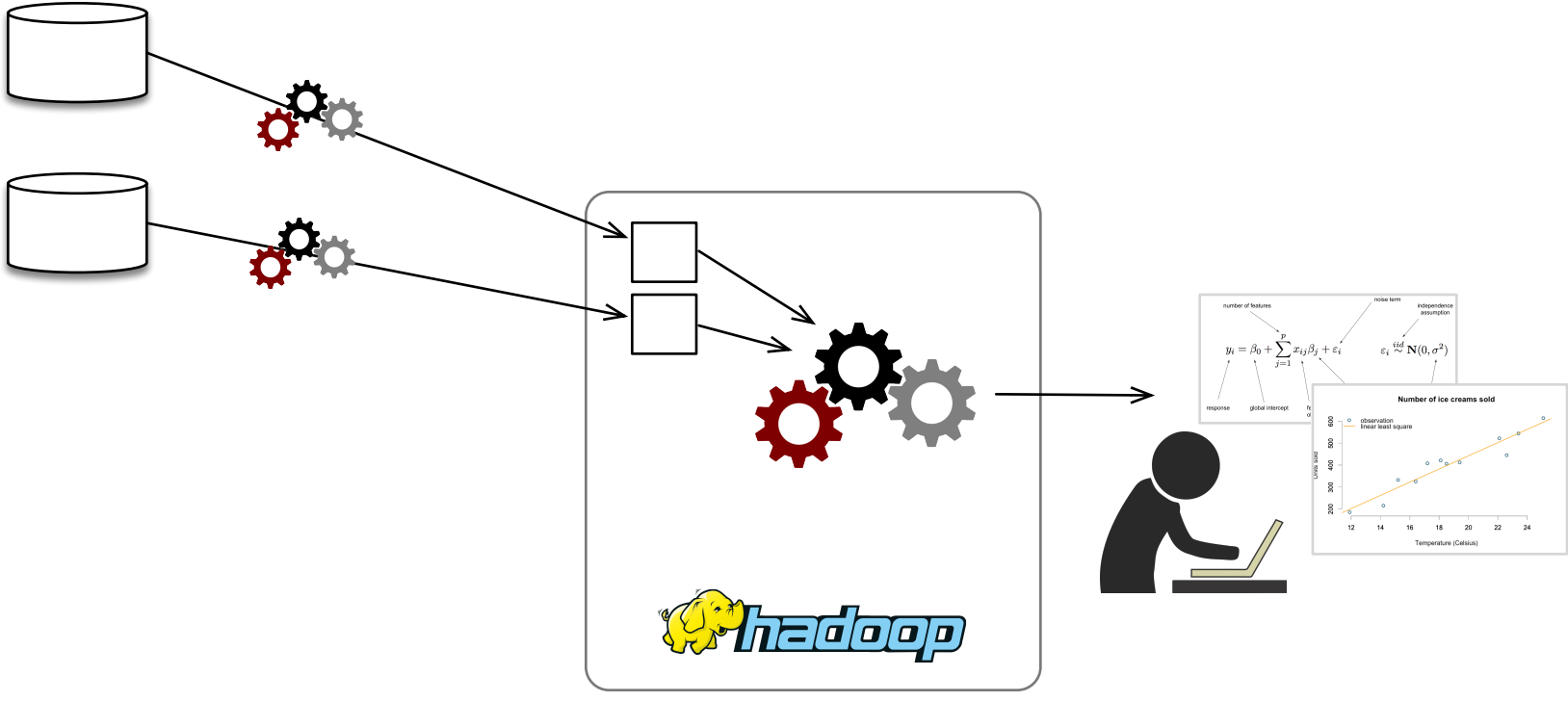


ETL Examples - IoT from set-top boxes



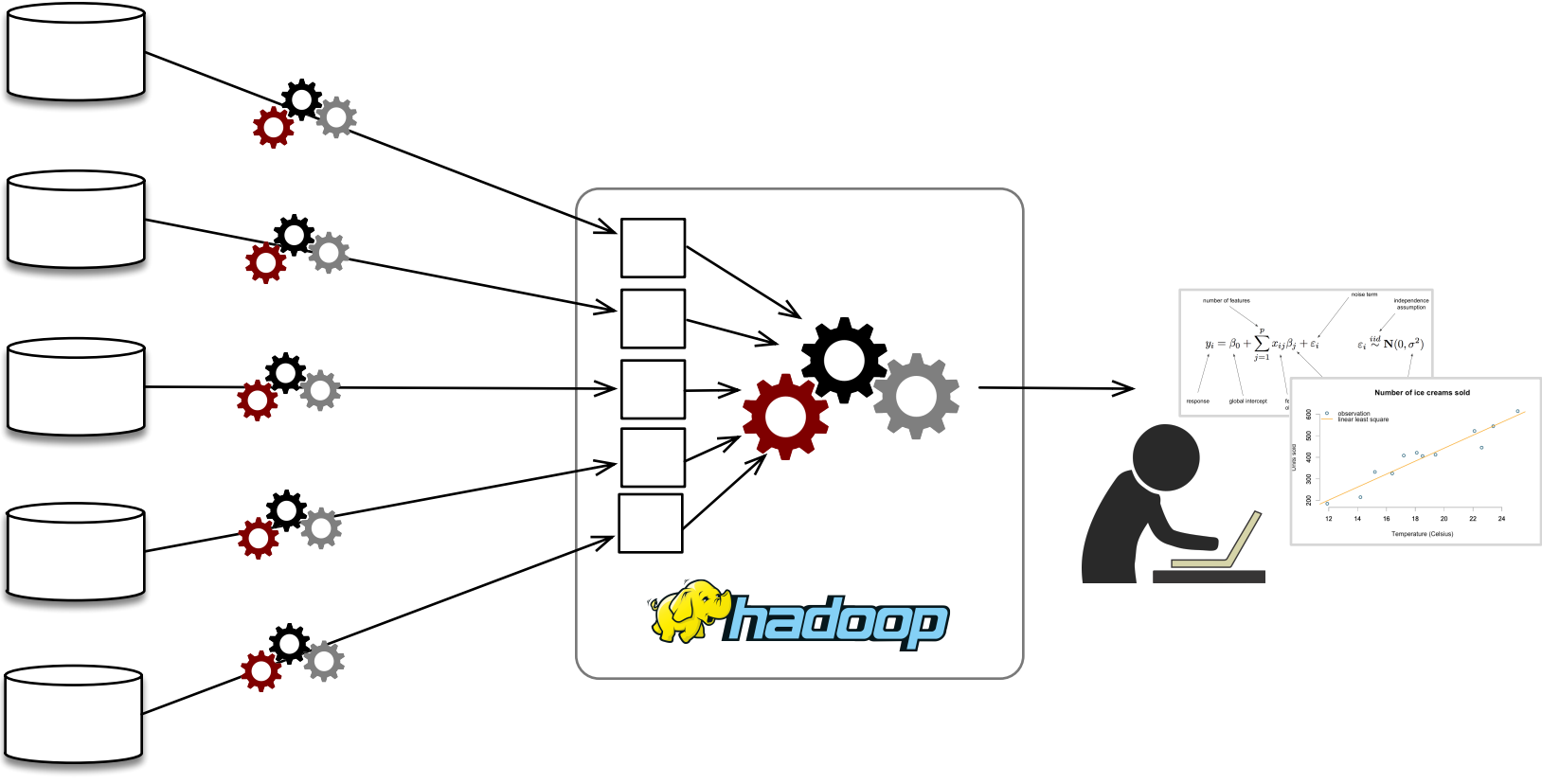
ETL Examples - data scientist pipelines

ETL Examples - data scientist pipelines



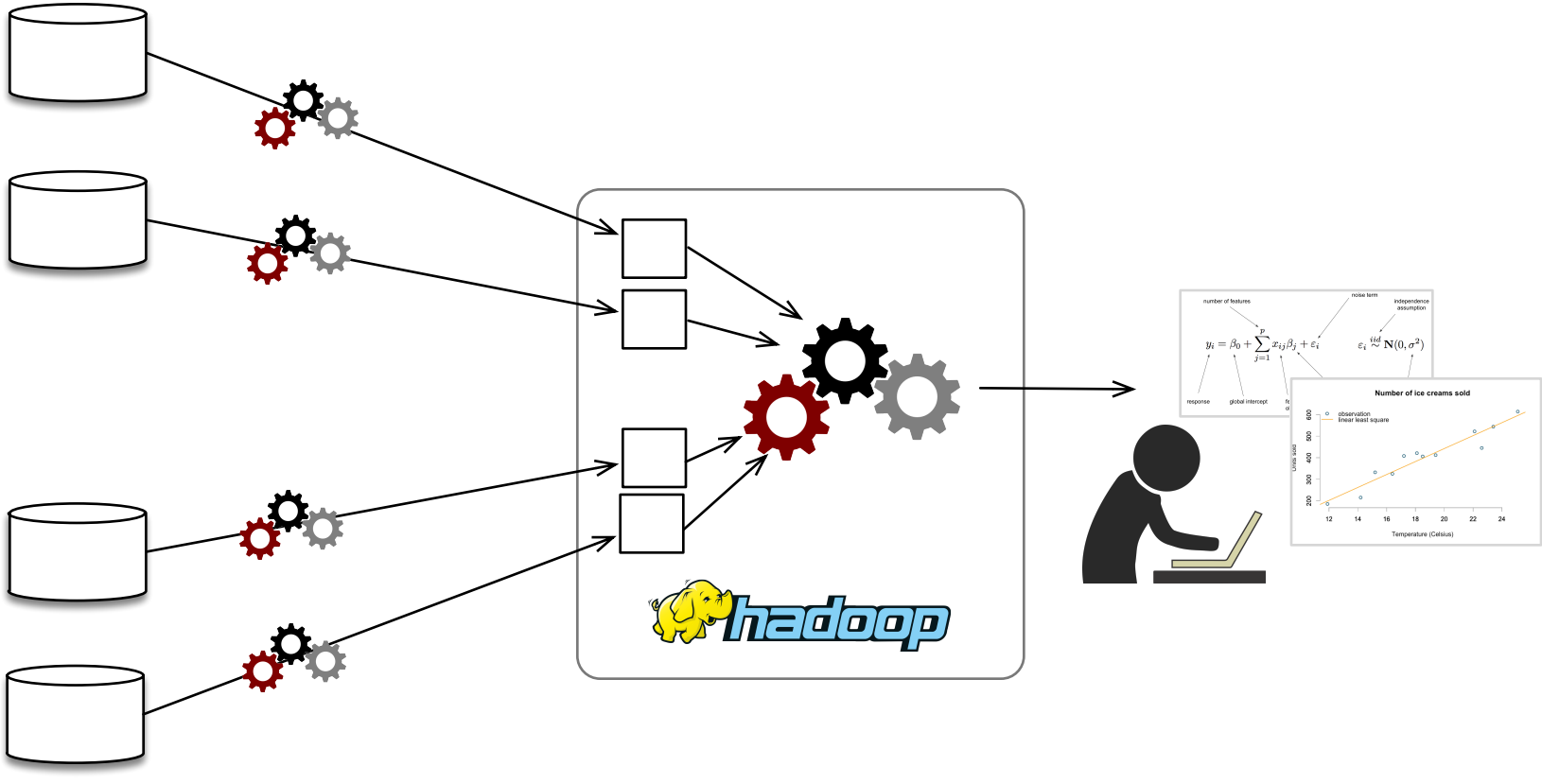
gluent.

ETL Examples - data scientist pipelines



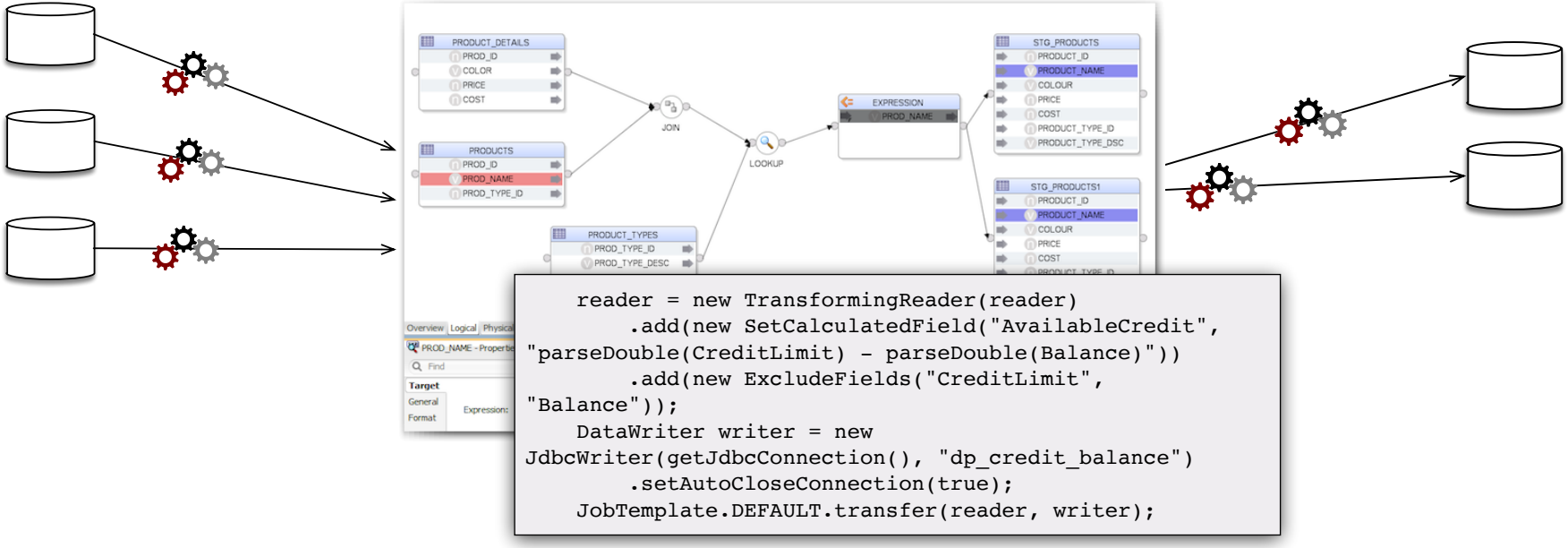
gluent.

ETL Examples - data scientist pipelines



gluent.

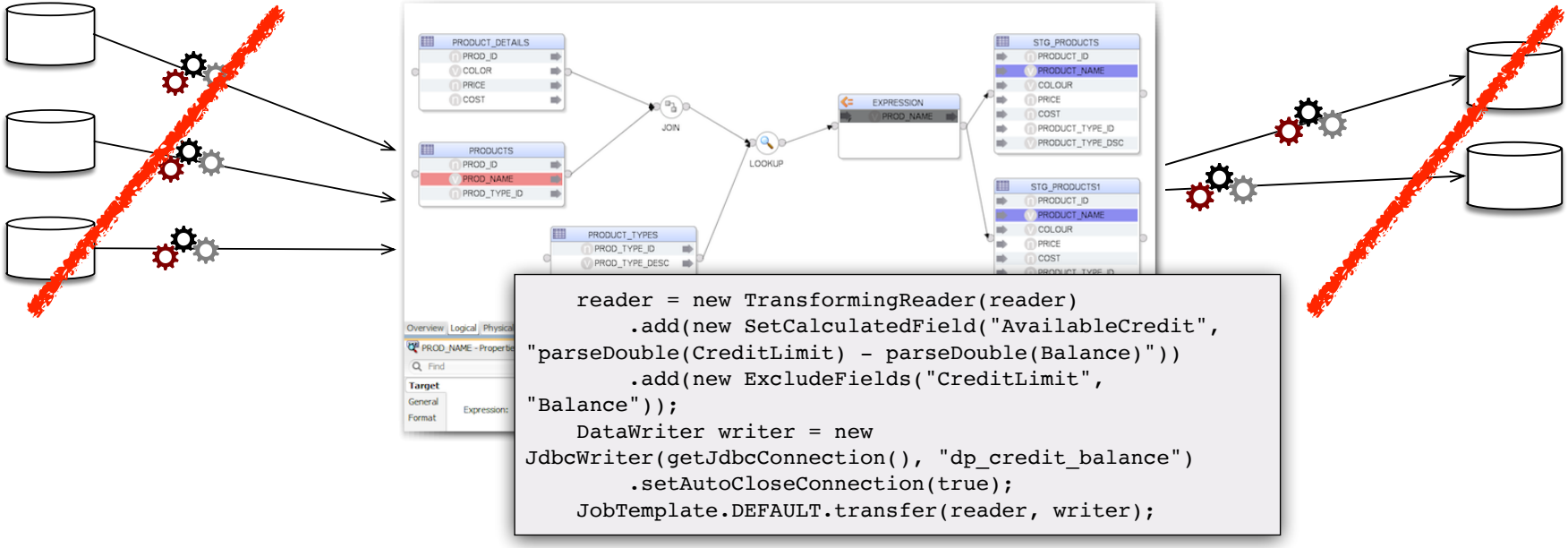
Typical ETL development - too much time spent on “E” and “L”




```
    reader = new TransformingReader(reader)
        .add(new SetCalculatedField("AvailableCredit",
"parseDouble(CreditLimit) - parseDouble(Balance)"))
        .add(new ExcludeFields("CreditLimit",
"Balance"));
    DataWriter writer = new
JdbcWriter(getJdbcConnection(), "dp_credit_balance")
        .setAutoCloseConnection(true);
    JobTemplate.DEFAULT.transfer(reader, writer);
```

gluent.

Typical ETL development - too much time spent on “E” and “L”




```
    reader = new TransformingReader(reader)
        .add(new SetCalculatedField("AvailableCredit",
"parseDouble(CreditLimit) - parseDouble(Balance)"))
        .add(new ExcludeFields("CreditLimit",
"Balance"));
    DataWriter writer = new
JdbcWriter(getJdbcConnection(), "dp_credit_balance")
        .setAutoCloseConnection(true);
    JobTemplate.DEFAULT.transfer(reader, writer);
```

ETL Developer / Data Engineer
should spend their time on

Transformations!

gluent.

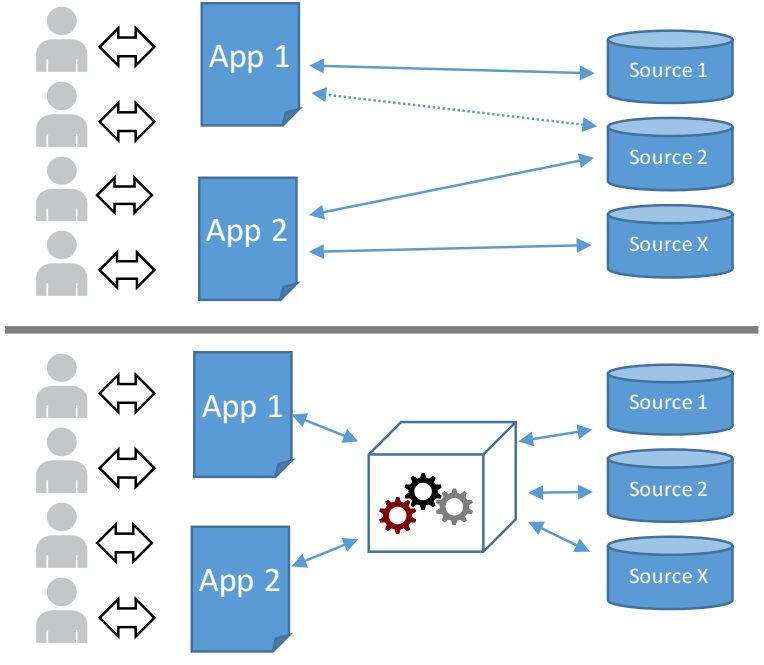
We still need to move data!

Data lake or data hub?

- Data lake - enterprise data stored in raw form
- Data hub - centralized data store with standards, governance, and quality

What about data federation?

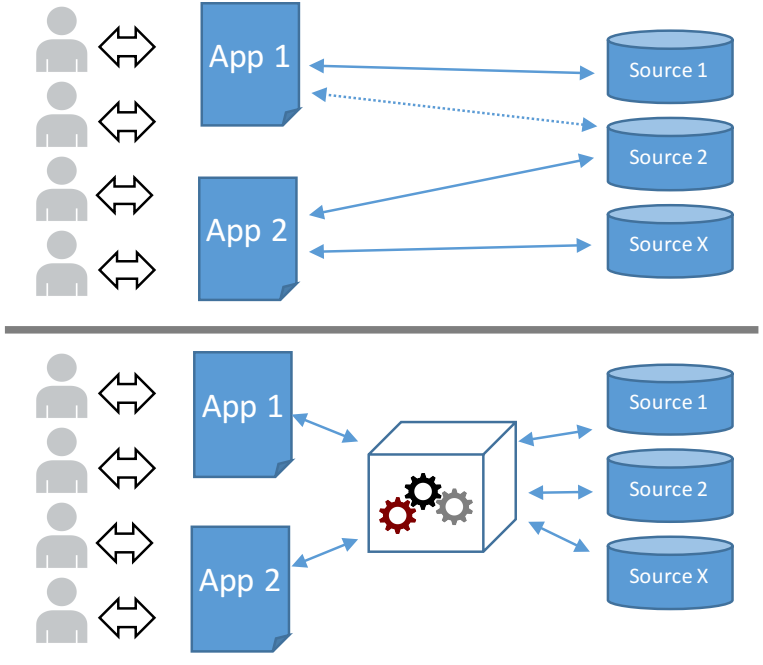
- Sharing data without moving data
- Metadata is stored about each “source” database
- Queries are passed through a metadata layer, which provides info about where the data lives and how to translate the query



What about data federation?

- Sharing data without moving data
- Metadata is stored about each “source” database
- Queries are passed through a metadata layer, which provides info about where the data lives and how to translate the query

- **But…**

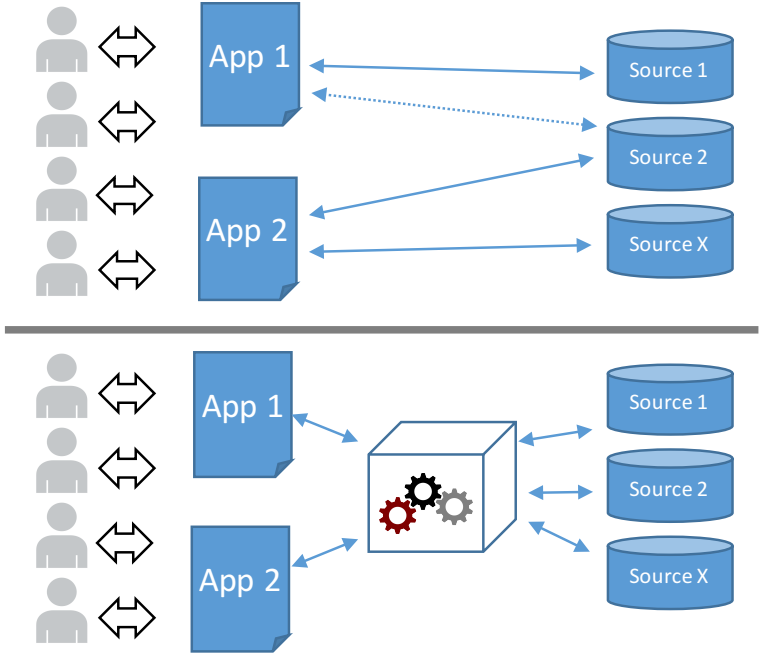


What about data federation?

- Sharing data without moving data
- Metadata is stored about each “source” database
- Queries are passed through a metadata layer, which provides info about where the data lives and how to translate the query

- **But…**

- We must recode applications!
- Data remains in RDBMS silos and are limited to CPU/storage constraints



gluent.

Three things we need the ability to do:

1. Offload data from Oracle to Hadoop
2. Load data from Hadoop to Oracle
3. Query Hadoop data in Oracle

Offload, not Extract

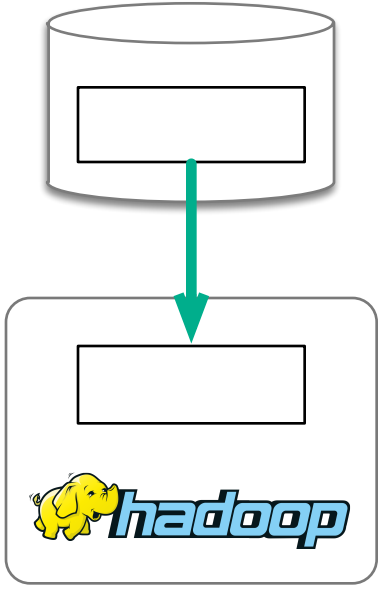
- "**Offload**" - copy or move data from the relational database to Hadoop

- Why offload?
  - Store data in a centralized location for access and sharing
  - Use the distributed, parallel processing power of Hadoop for transformations
  - Enable the use of "new world" technologies (Spark, Impala, etc)

- Why Hadoop?
  - We can now afford to keep a copy of all enterprise data for data sharing reasons!

How to offload? There are many options

Tool	Offload Data
Sqoop	Yes
Oracle Loader for Hadoop	
Oracle SQL Connector for HDFS	
ODBC Gateway	
Big Data SQL	
Gluent Data Platform	Yes



Bulk data load with Sqoop

- Command line client used to bulk copy data from a relational database to HDFS over JDBC connection
  - Also works in reverse, HDFS to RDBMS
  - Sqoop generates MapReduce jobs for the work

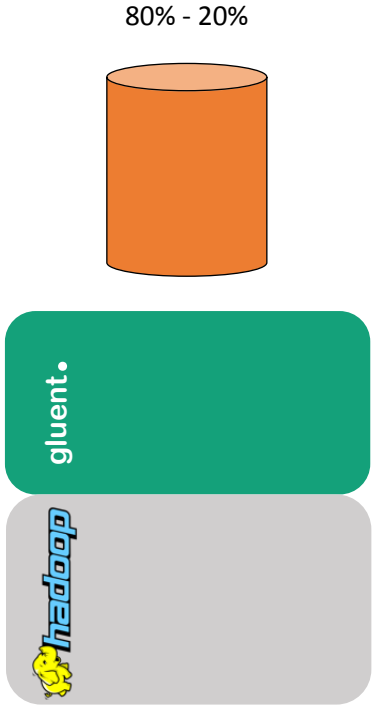
```
sqoop import --connect jdbc:oracle:thin:@myserver:1521/MYDB1
   --username myuser
   --null-string ''
   --null-non-string ''
   --target-dir=/user/hive/warehouse/ssh/sales
   --append -m1 --fetch-size=5000
   --fields-terminated-by '','' --lines-terminated-by ''\\n''
   --optionally-enclosed-by ''\"'' --escaped-by ''\"''
   --split-by TIME_ID
   --query "\"SELECT * FROM SSH.SALES WHERE TIME_ID < DATE'1998-01-01'\""
```

Offload data with Gluent Data Platform

- Gluent Advisor helps determine which tables and/or partitions are candidates for offload
- Offload options:
  - Move or copy 100% of data
  - Move “cold” partitions only
- Data can be kept in-sync with an incremental offload
  - Example: when a partition goes inactive (cold), it can be offloaded to Hadoop
- Data can be updated from RDBMS to Hadoop

Offload data with Gluent Data Platform

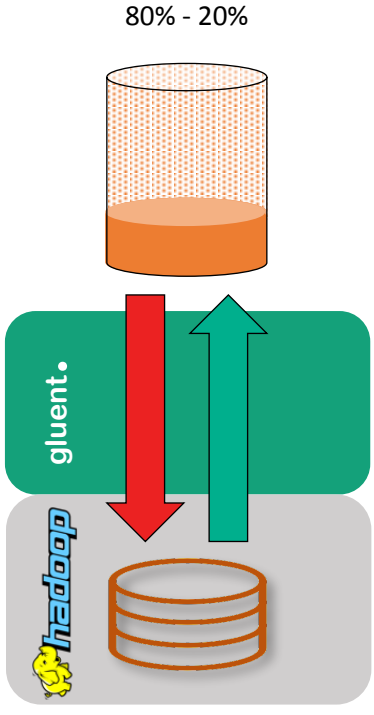
- Gluent Advisor helps determine which tables and/or partitions are candidates for offload
- Offload options:
  - Move or copy 100% of data
  - Move “cold” partitions only
- Data can be kept in-sync with an incremental offload
  - Example: when a partition goes inactive (cold), it can be offloaded to Hadoop
- Data can be updated from RDBMS to Hadoop




```
offload -x -t EDW.BALANCE_DETAIL_MONTHLY --less-than-value=2017-05-01
```

Offload data with Gluent Data Platform

- Gluent Advisor helps determine which tables and/or partitions are candidates for offload
- Offload options:
  - Move or copy 100% of data
  - Move “cold” partitions only
- Data can be kept in-sync with an incremental offload
  - Example: when a partition goes inactive (cold), it can be offloaded to Hadoop
- Data can be updated from RDBMS to Hadoop




```
offload -x -t EDW.BALANCE_DETAIL_MONTHLY --less-than-value=2017-05-01
```

Three things we need the ability to do:

1. Offload data from Oracle to Hadoop
2. Load data from Hadoop to Oracle
3. Query Hadoop data in Oracle

Three things we need the ability to do:

1. Offload data from Oracle to Hadoop

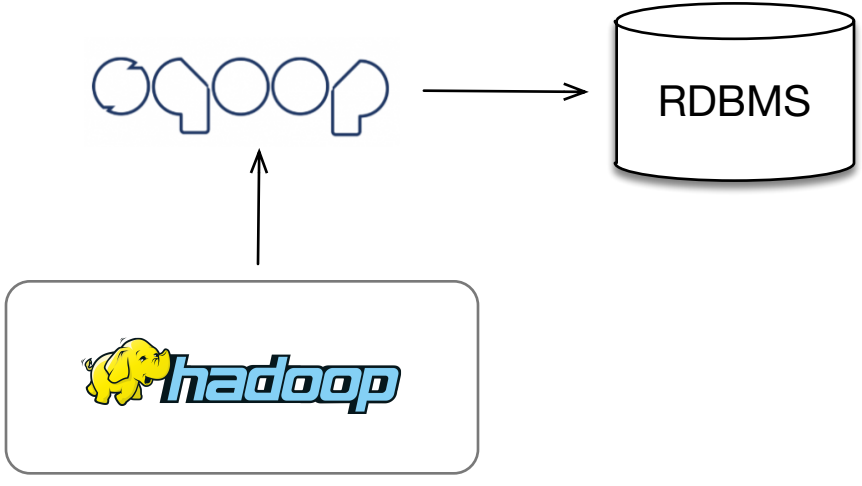
Present

2. Load data from Hadoop to Oracle

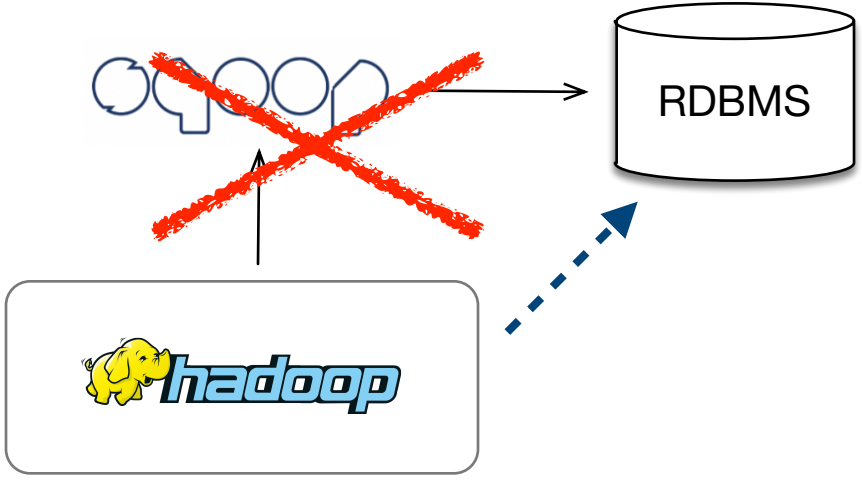
3. Query Hadoop data in Oracle

The “L” of ETL - Loading the data

- Sqoop in reverse
  - Reads HDFS files, converts to JDBC arrays and insert (or update) statements

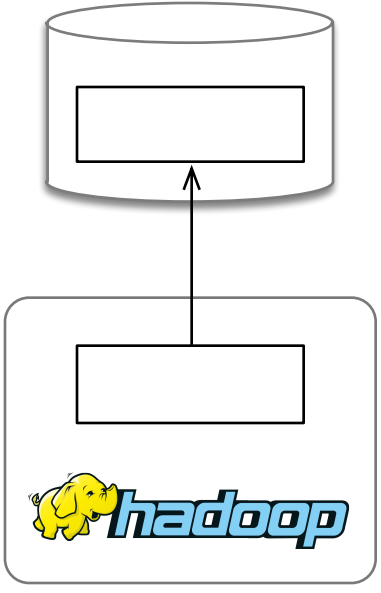


Forget the “L”oad. Present the data to the RDBMS



Present

Load the data from Hadoop to RDBMS?

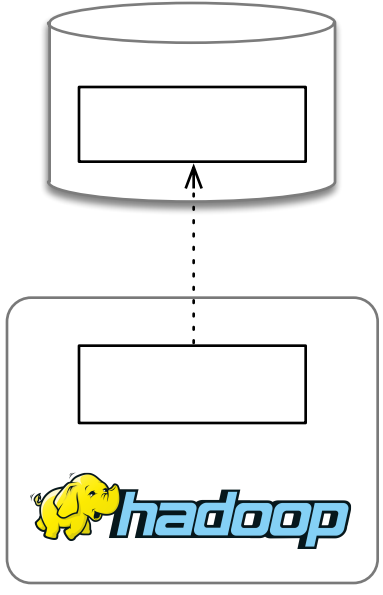


Present

Load the data from Hadoop to RDBMS?

Tool	Offload Data	Load Data
Sqoop	Yes	Yes
Oracle Loader for Hadoop		Yes
Oracle SQL Connector for HDFS		Yes
ODBC Gateway		Yes
Big Data SQL		Yes
Gluent Data Platform	Yes	Yes

Present



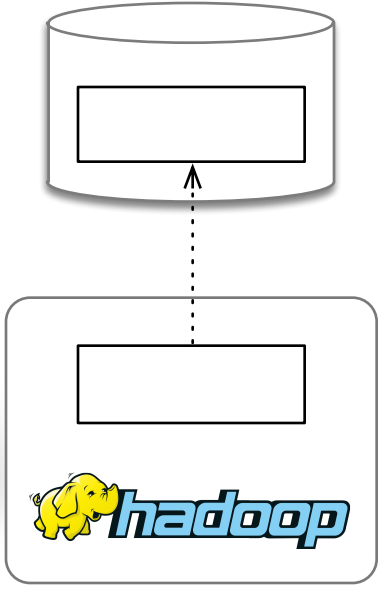
gluent.

Present

Load the data from Hadoop to RDBMS?

Tool	Offload Data	Load Data
Sqoop	Yes	Yes
Oracle Loader for Hadoop		Yes
Oracle SQL Connector for HDFS		
ODBC Gateway		
Big Data SQL		
Gluent Data Platform		

Present

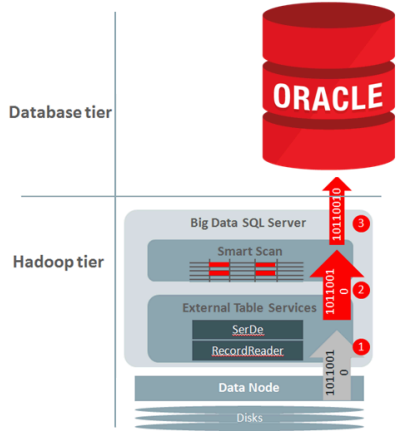


Present data to Oracle

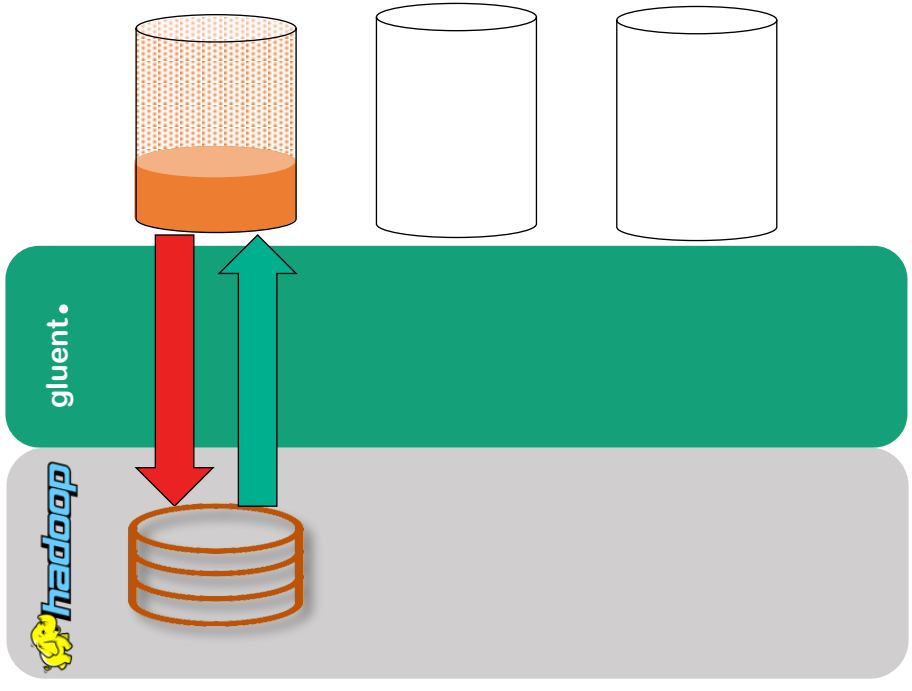
- Oracle SQL Connector for HDFS
  - Uses Oracle External Table to access delimited file or Hive table in Hadoop
  - When queried, all data is read from Hadoop to Oracle, then processed

- Oracle to Hadoop over database links
  - Create a database link from Oracle to Hadoop using the Impala (or Hive) ODBC gateway
  - Pushes filters to Hadoop but not grouping aggregates

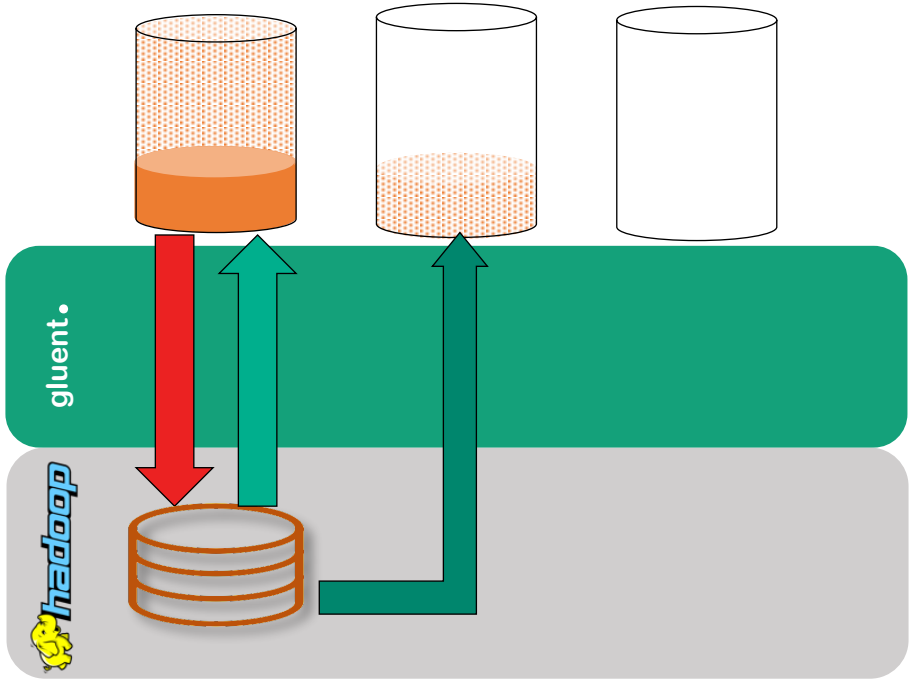
- Big Data SQL
  - Access Hive tables via an Oracle external table
  - Predicate pushdown (not aggregations & joins)



Gluent Present

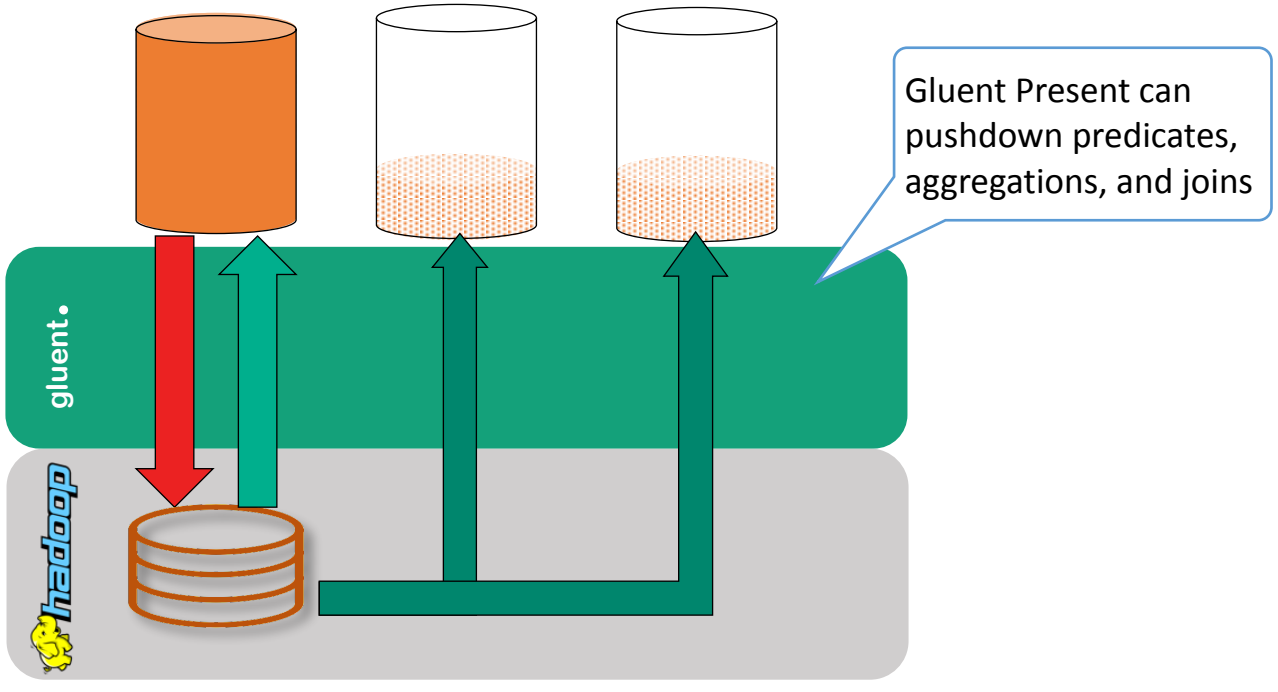


Gluent Present




```
present -t "SH.SALES" -xv --target-name="SH.SALES_DIVISION_X"
```

Gluent Present




```
present -t "SH.SALES" -xv --target-name="SH.SALES_DIVISION_X"
```

Three things we need the ability to do:

1. Offload data from Oracle to Hadoop
2. Load data from Hadoop to Oracle

Present

3. Query Hadoop data in Oracle

How to query the data? There are many options

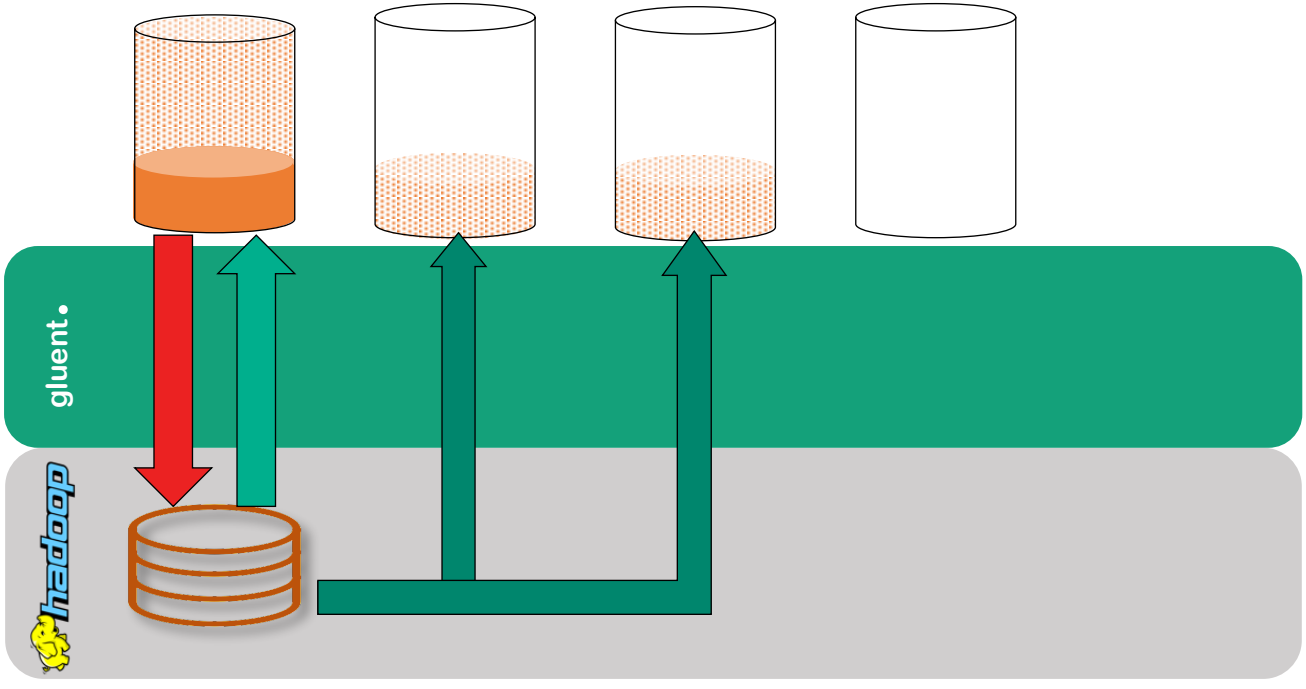
Tool	Offload Data	Load Data	Allow Query Data	Offload Query	Parallel Execution
Sqoop	Yes	Yes			Yes
Oracle Loader for Hadoop		Yes			Yes
Oracle SQL Connector for HDFS		Yes	Yes		Yes
ODBC Gateway		Yes	Yes	Yes	
Big Data SQL			Yes	Yes	Yes
Gluent Data Platform	Yes		Yes	Yes	Yes

How to query the data? There are many options

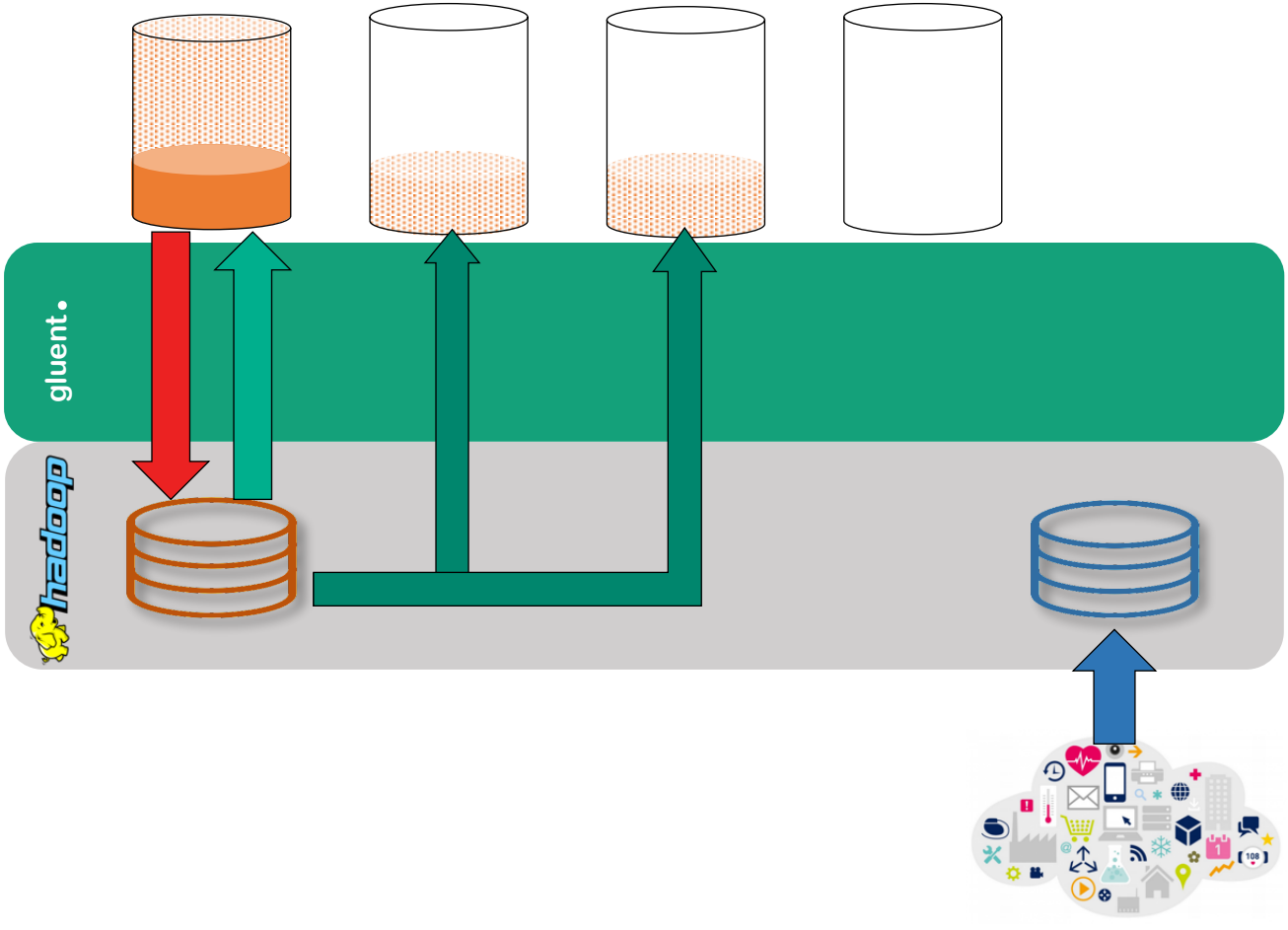
Tool	Offload Data	Load Data	Allow Query Data	Offload Query	Parallel Execution
Sqoop	Yes	Yes			Yes
Oracle Loader for Hadoop		Yes			Yes
Oracle SQL Connector for HDFS		Yes	Yes		Yes
ODBC Gateway		Yes	Yes	Yes	
Big Data SQL			Yes	Yes	Yes
Gluent Data Platform	Yes		Yes	Yes	Yes

Present

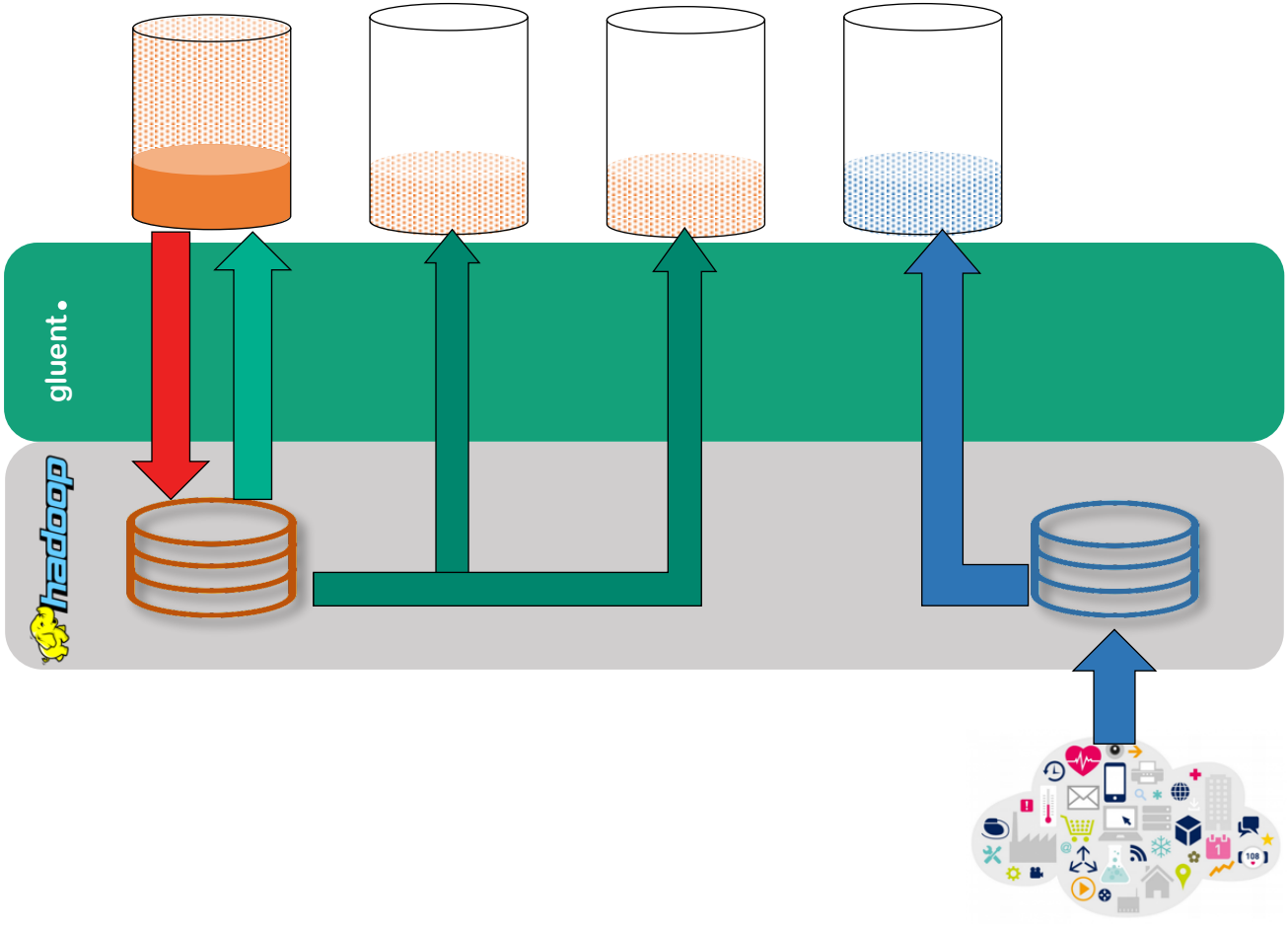
Present Data From Anywhere To Anywhere



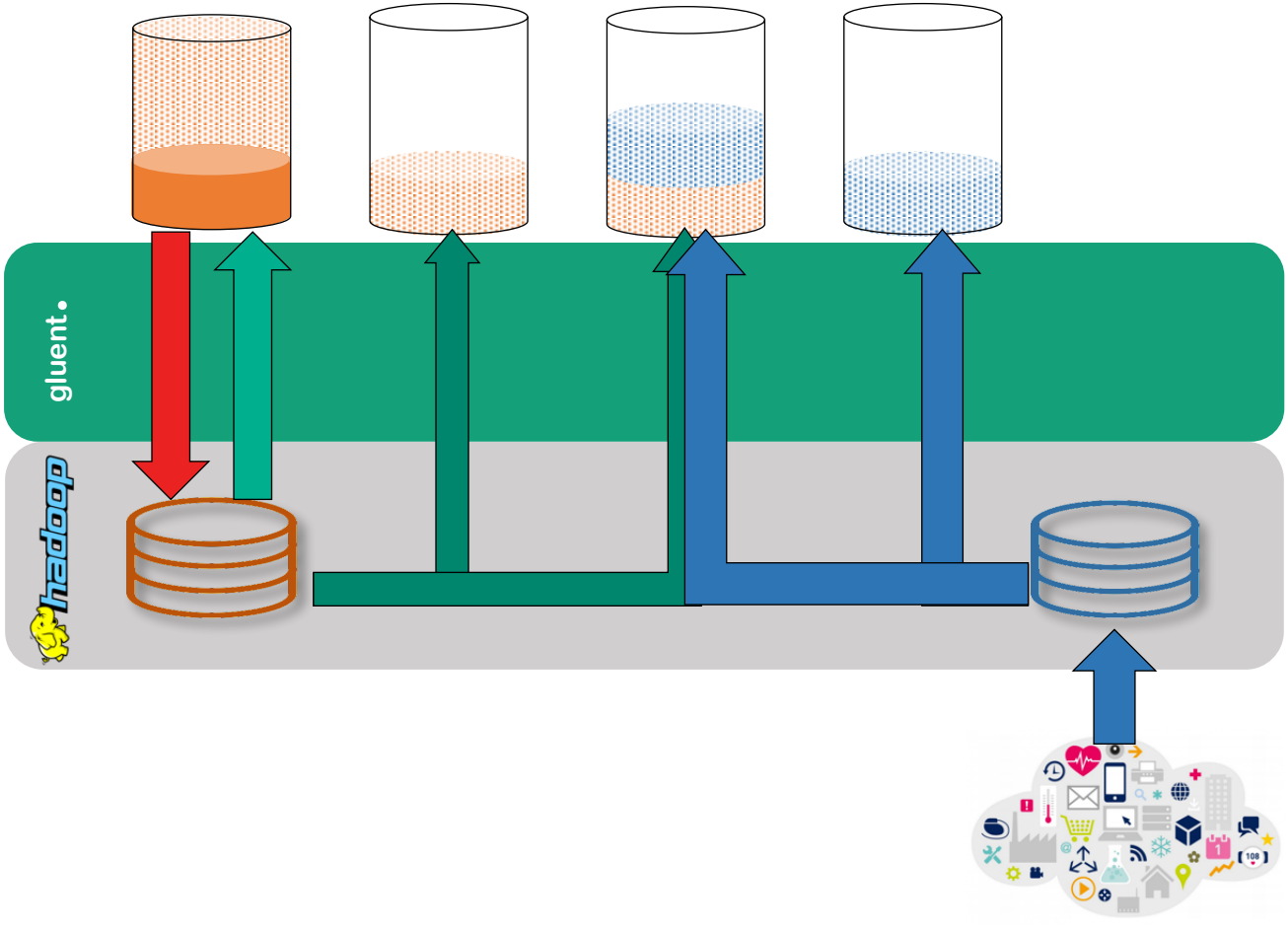
Present Data From Anywhere To Anywhere



Present Data From Anywhere To Anywhere

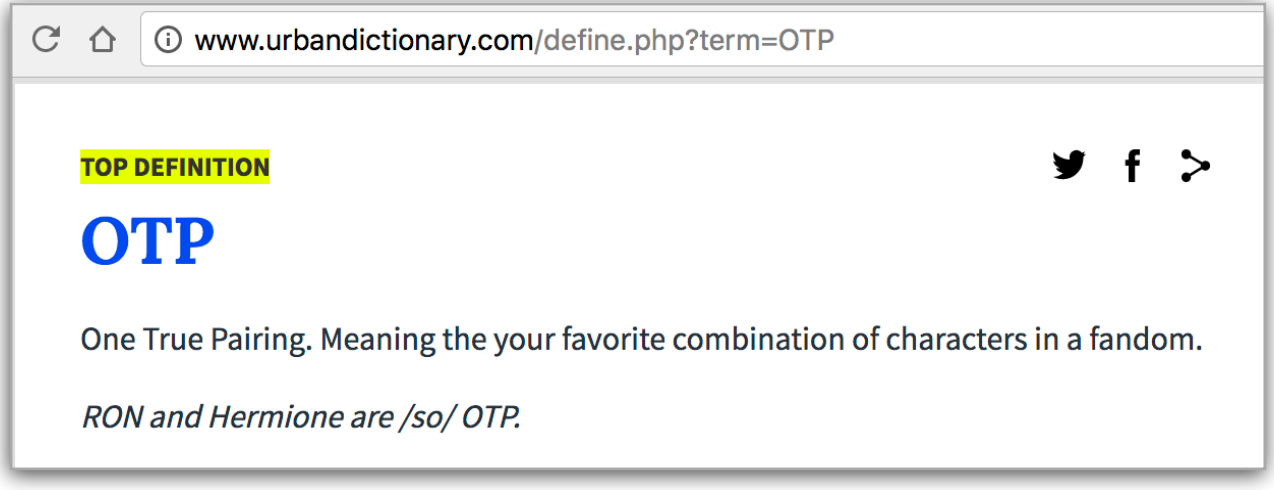


Present Data From Anywhere To Anywhere

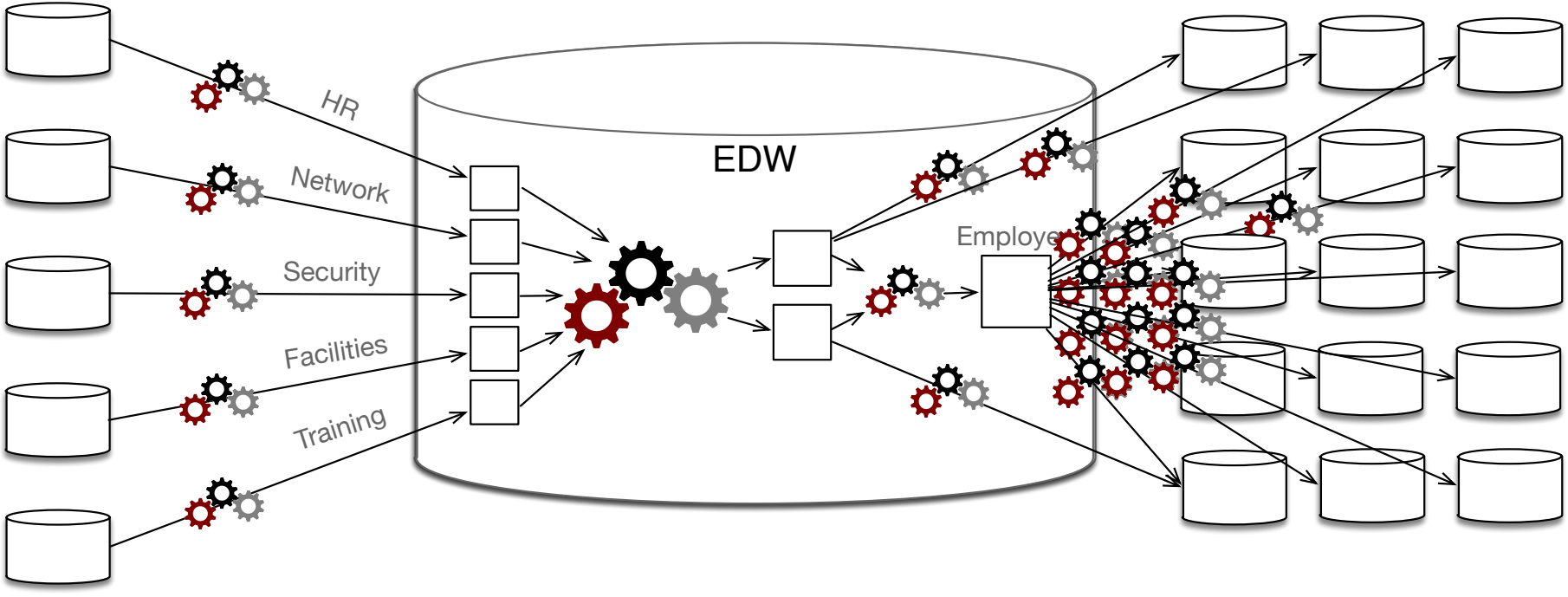


Offload Transform Present (OTP?)

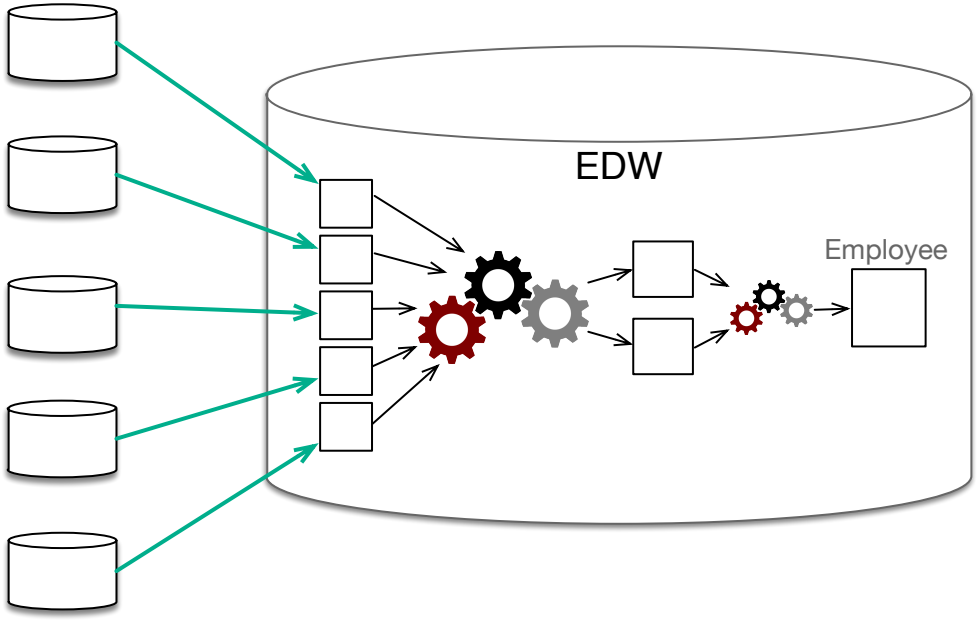
- A new approach…a new acronym?



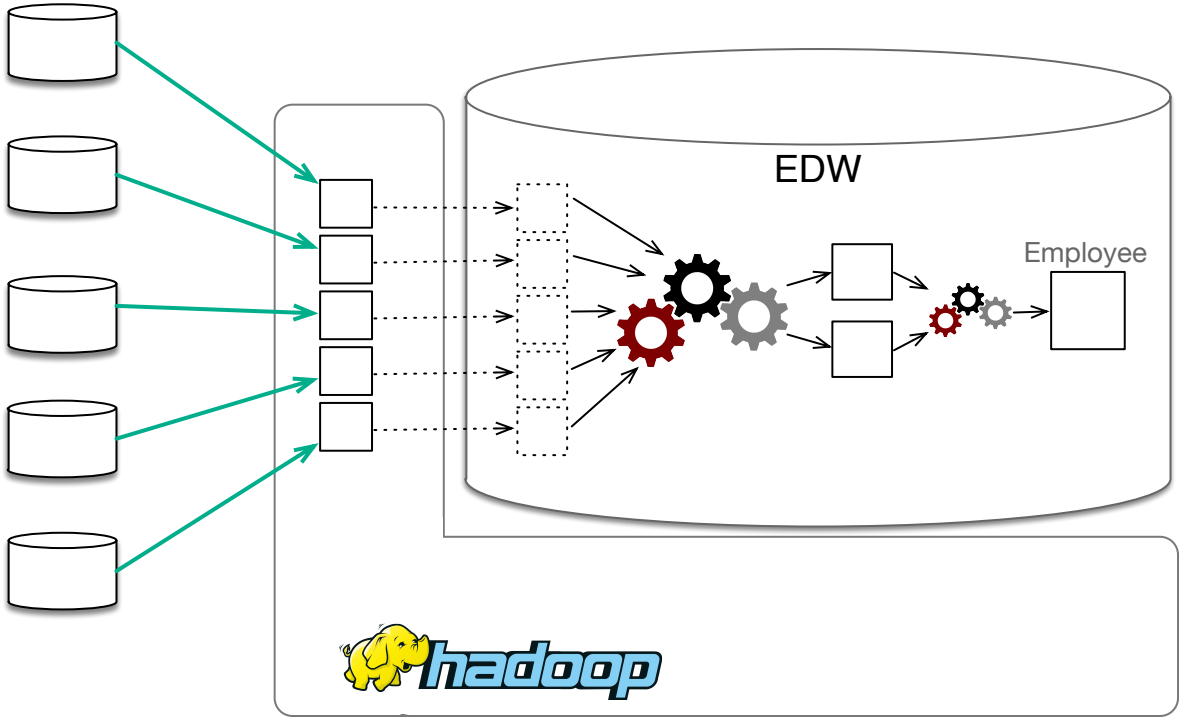
ETL Examples - employee data



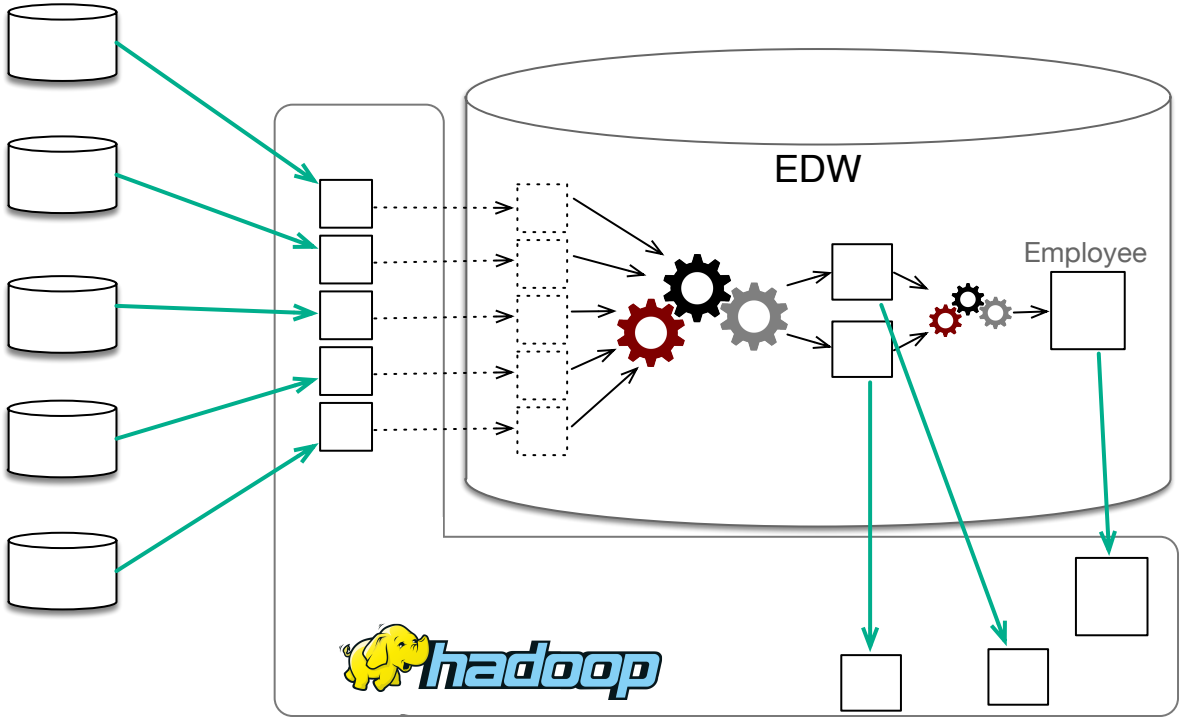
Now using OTP - employee data



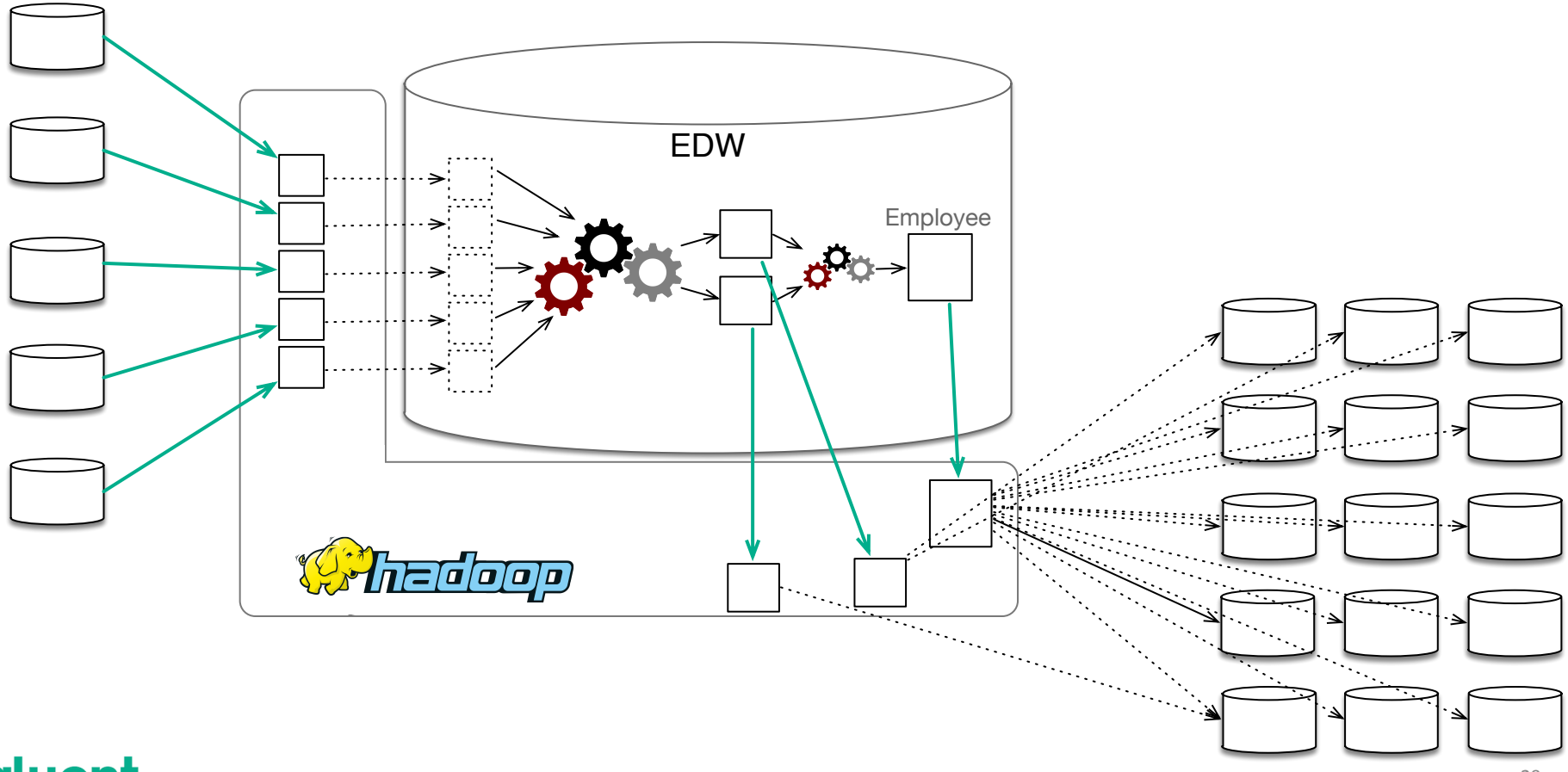
Now using OTP - employee data



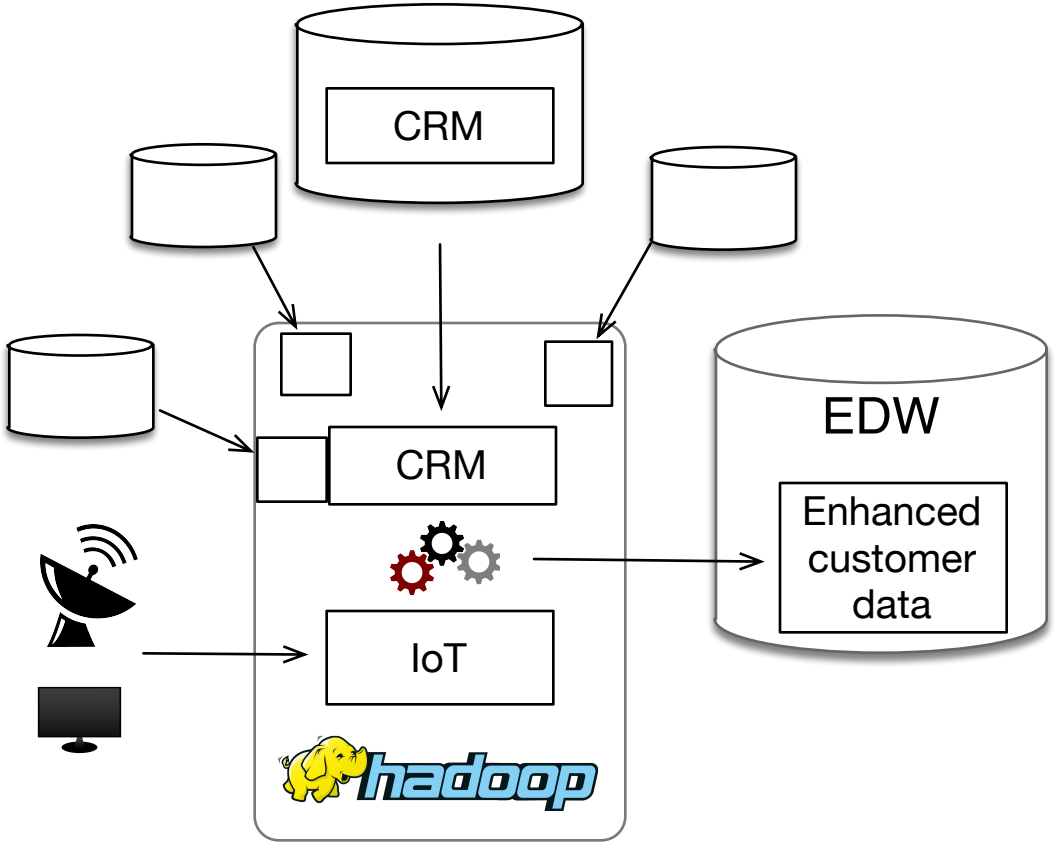
Now using OTP - employee data



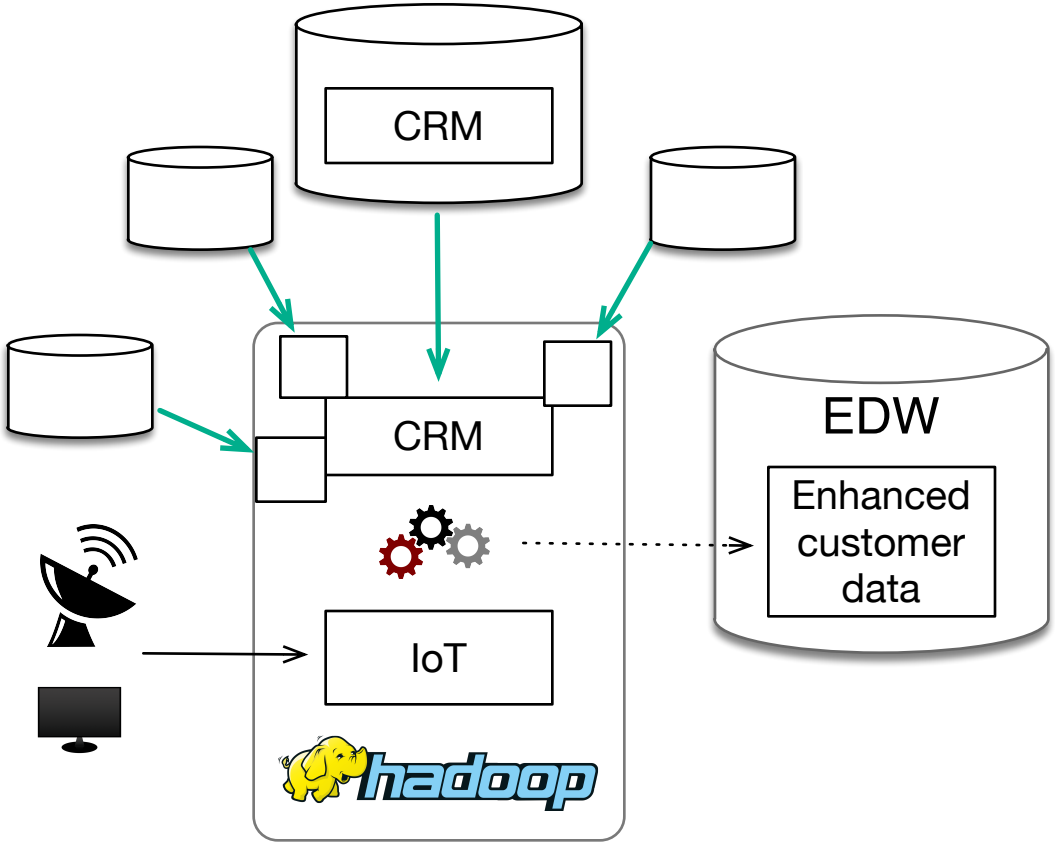
Now using OTP - employee data



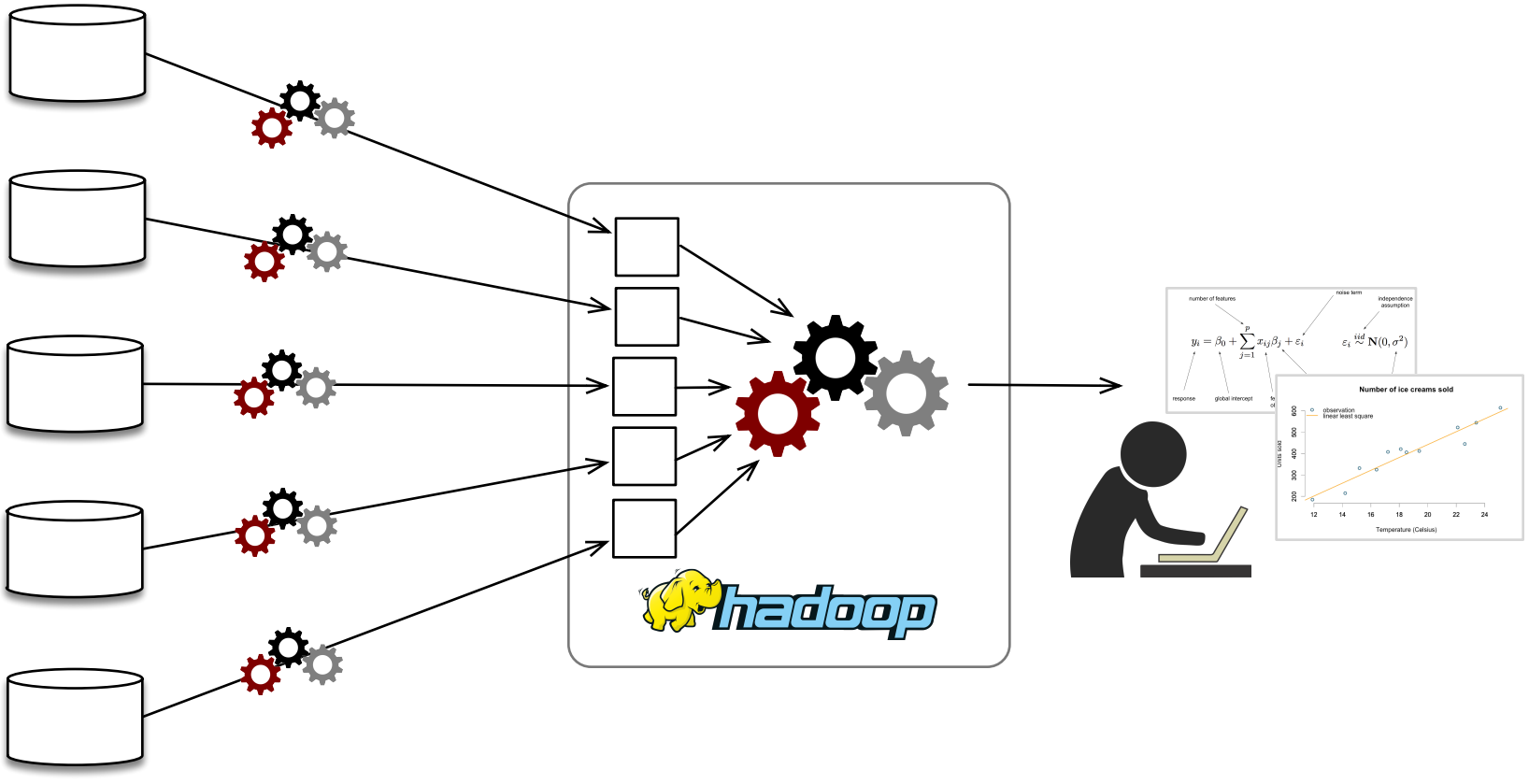
ETL Examples - IoT from set-top boxes



ETL Examples - IoT from set-top boxes

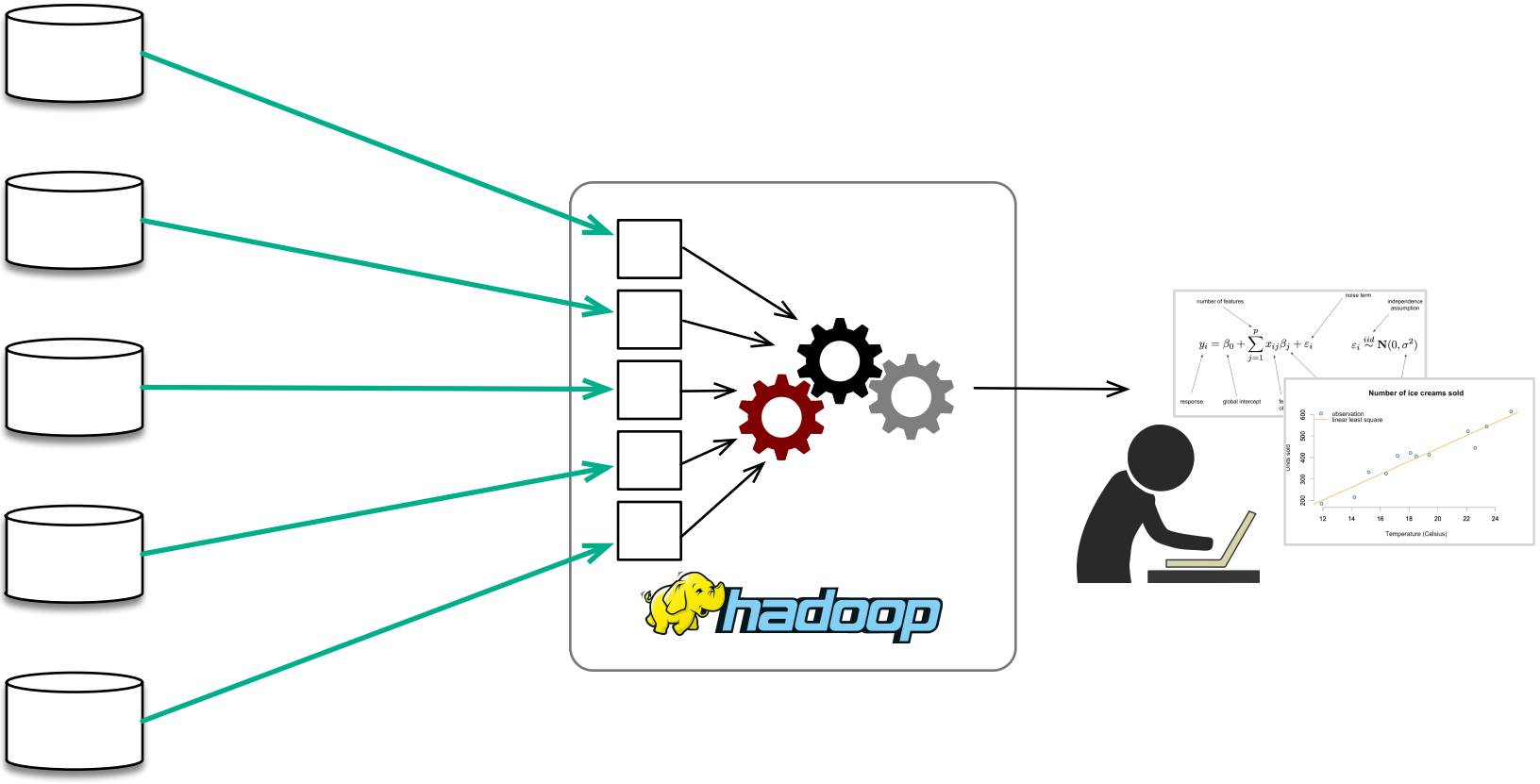


ETL Examples - data scientist pipelines



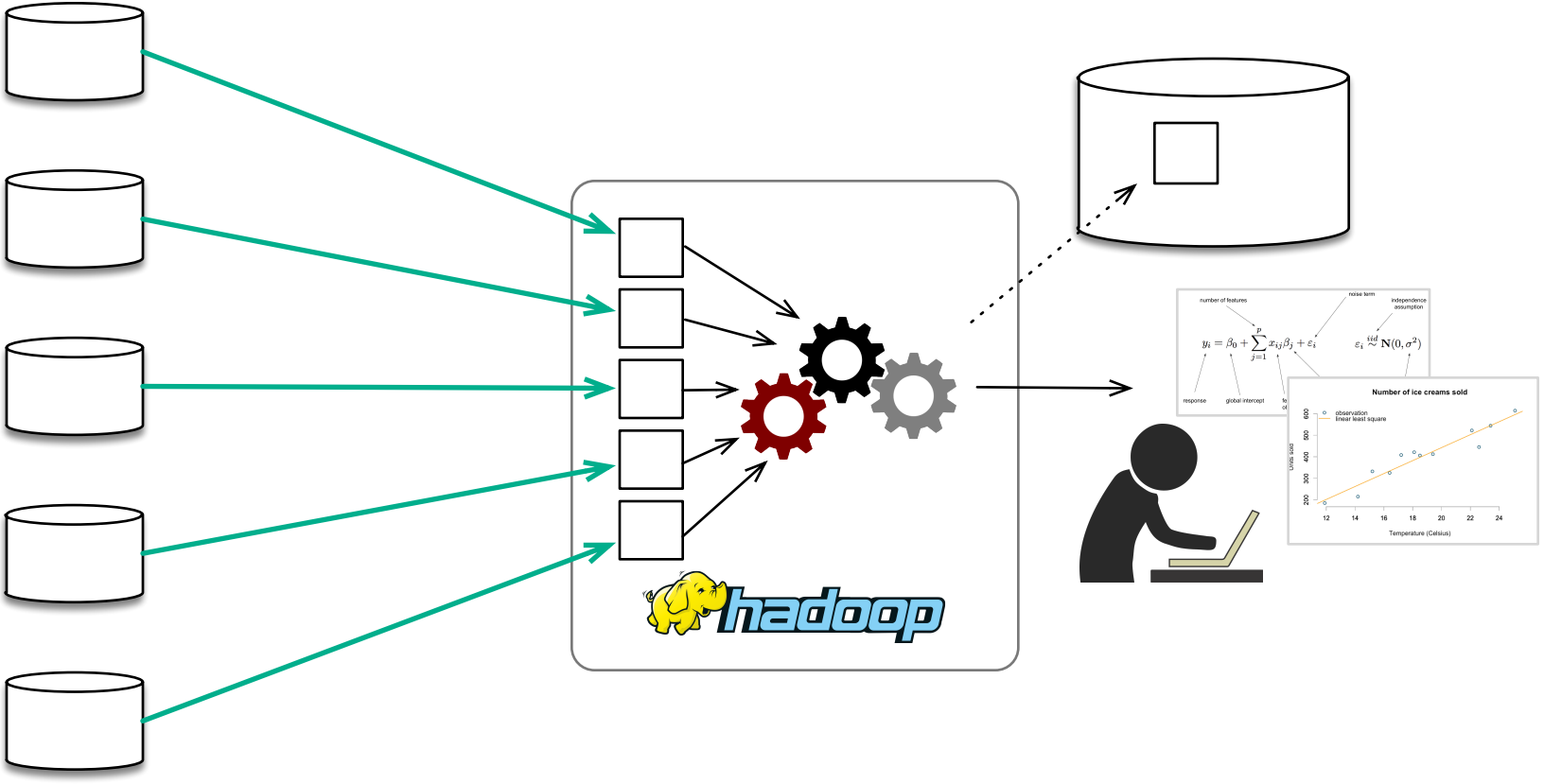
gluent.

ETL Examples - data scientist pipelines



gluent.

ETL Examples - data scientist pipelines



gluent.

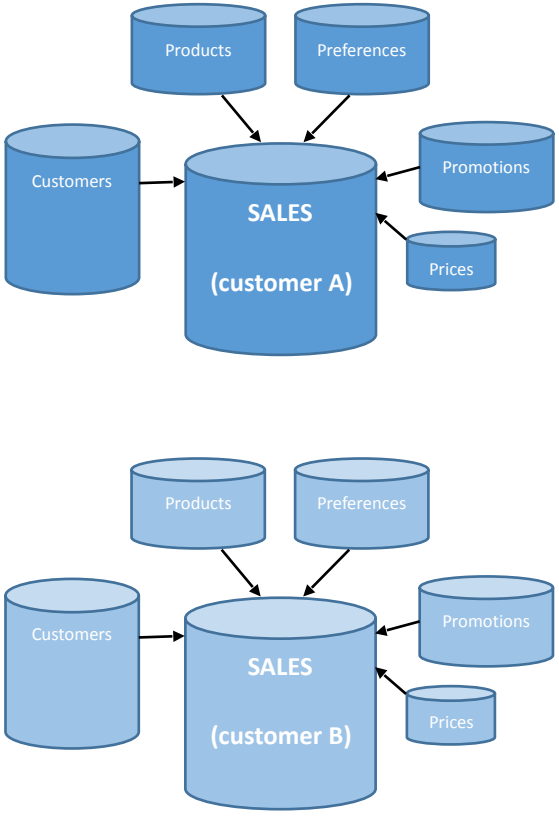
OTP simplifies the process of data movement!

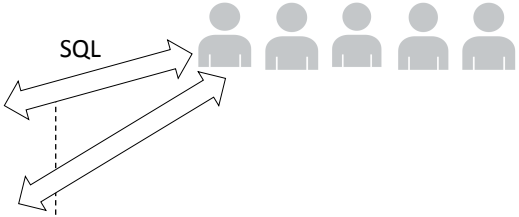
- No upfront massive project cost to move data around.
  - Can happen in minutes / hours, not weeks / months
  - Large corporations can move as fast as startups

- Data Engineers / ETL Developers can focus on **transformations**!

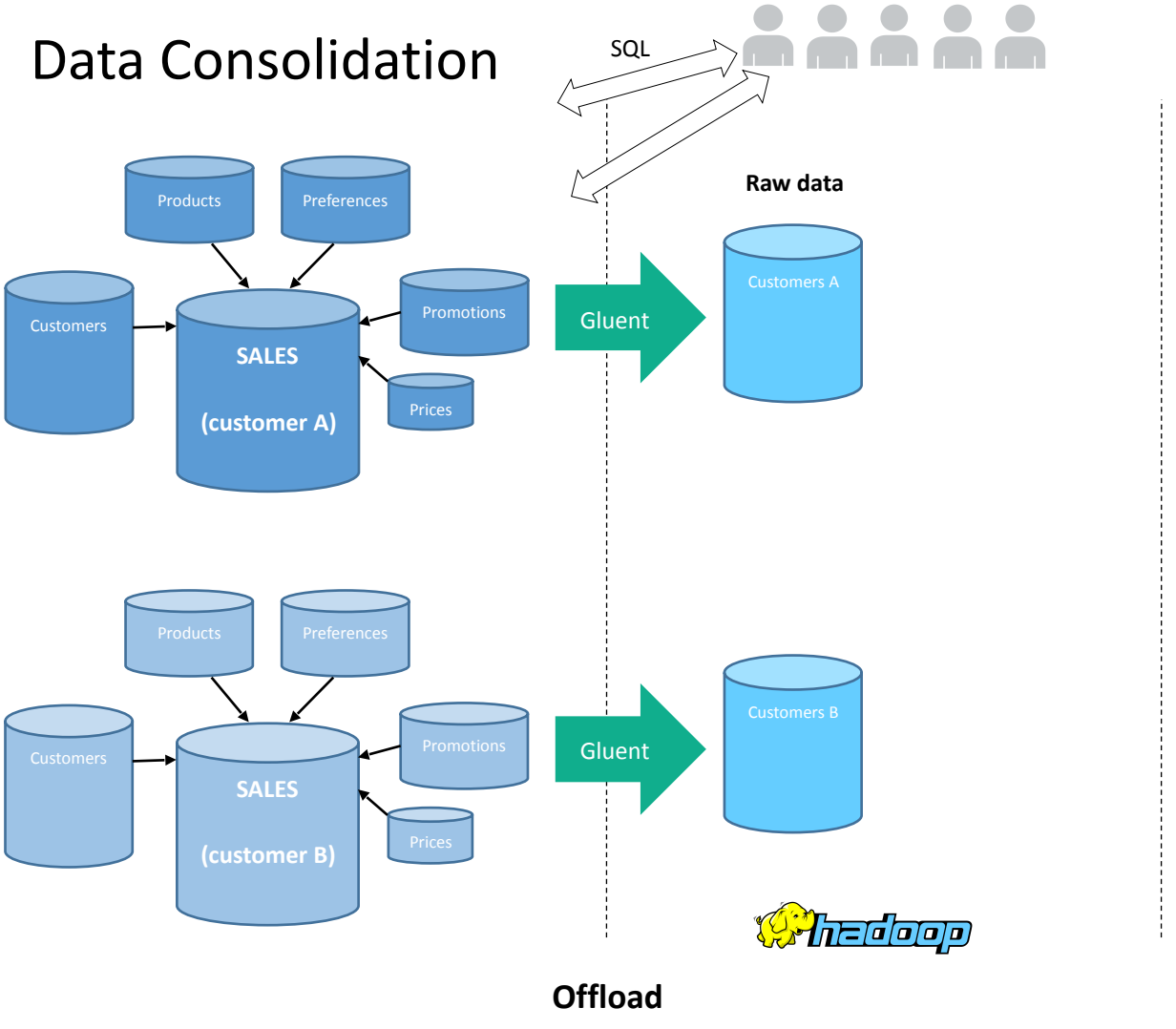
- A new approach to data sharing across enterprise
  - Offload ***all*** data for simple access
  - Takes a different mindset

Data Consolidation

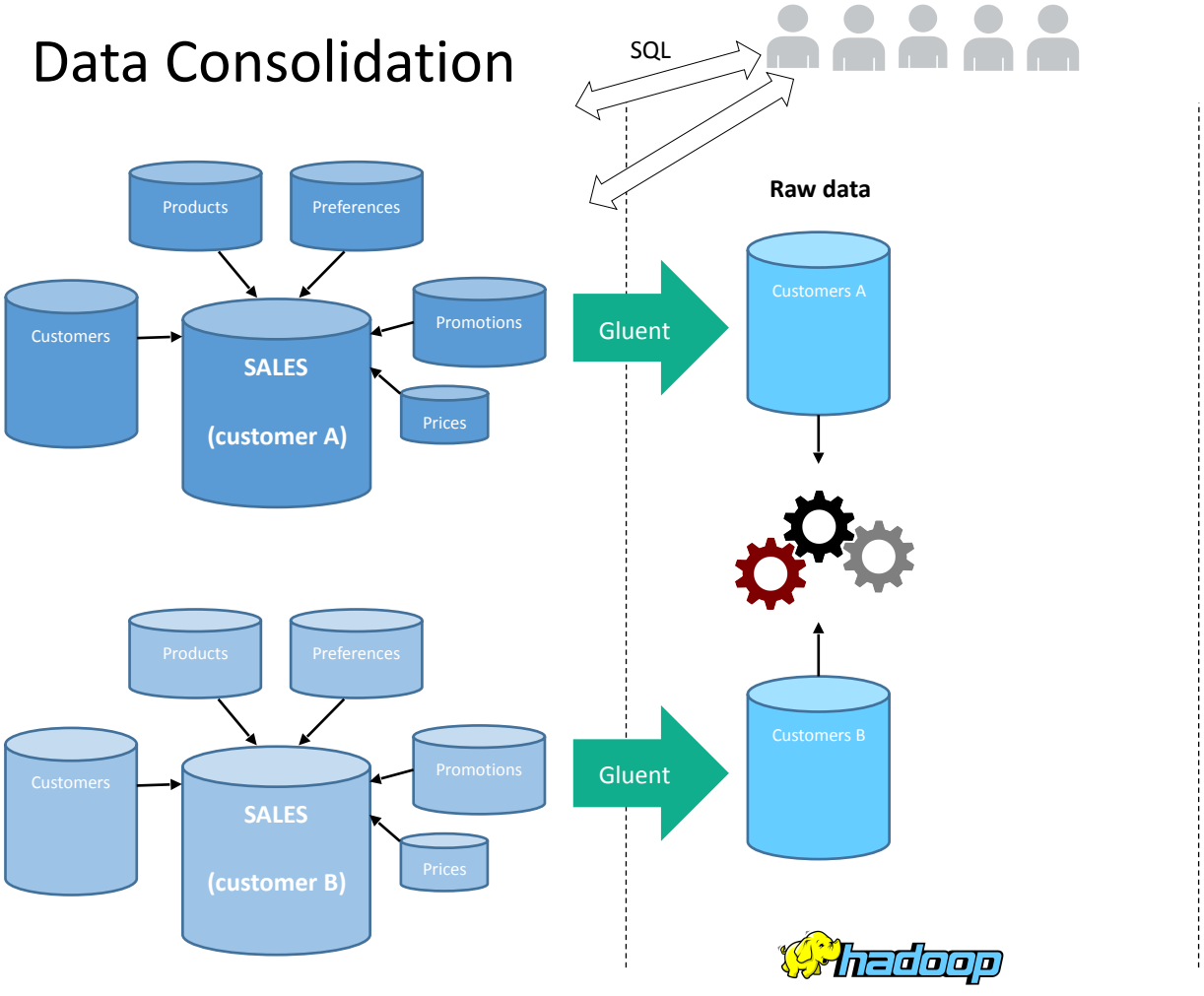




Data Consolidation

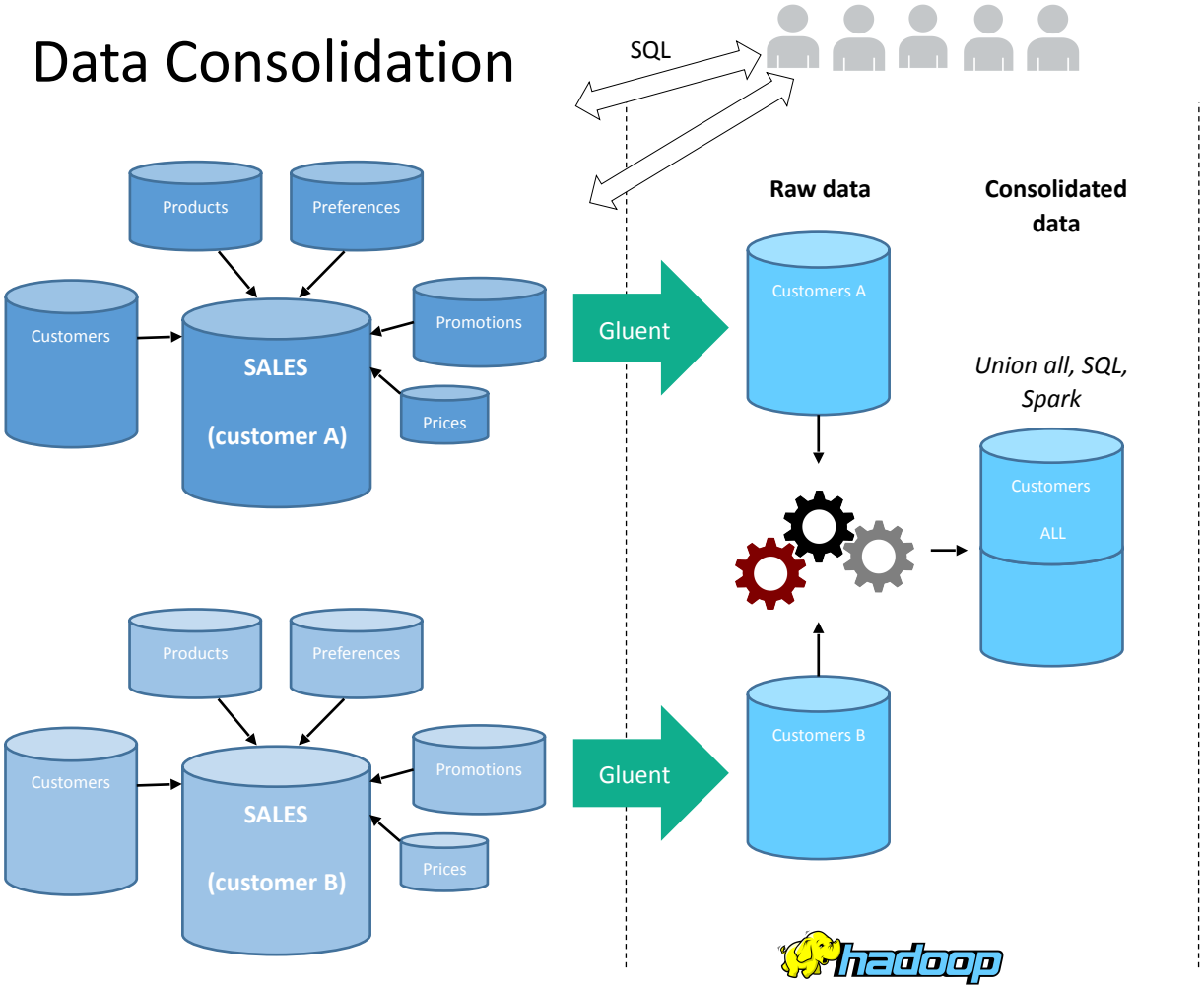


Data Consolidation



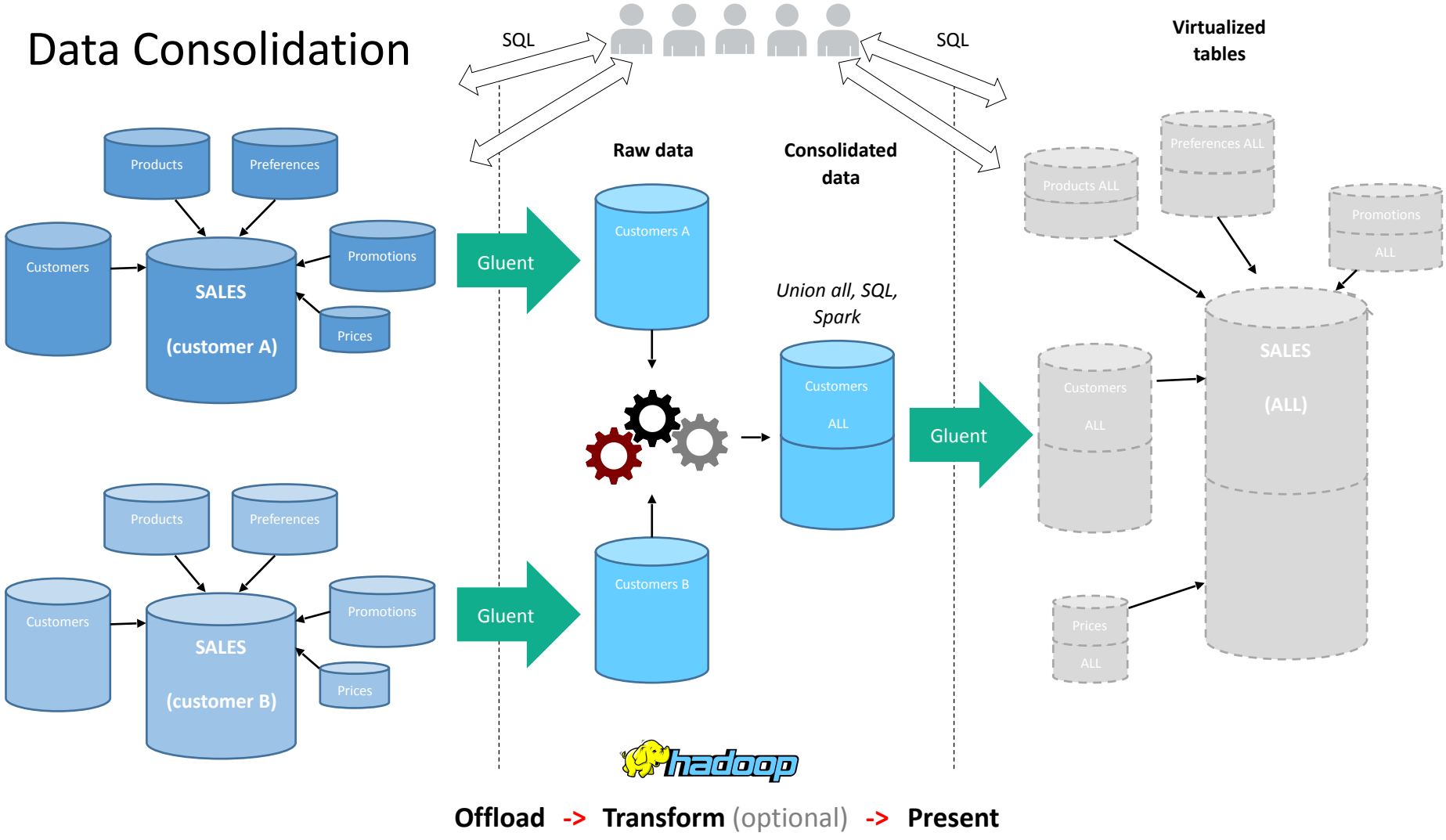
Offload -> Transform (optional)

Data Consolidation



Offload -> Transform (optional)

Data Consolidation



Offload -> **Transform** (optional) -> **Present**

thank you!

Gluent Webinars - JULY 2017

Gluent is running one-hour webinars each Wednesday in July beginning tomorrow, July 12. See details below and register for your free spot today!

Apache Impala Internals

Speaker: Tanel Poder, Gluent

Wednesday, July 19 @ 12 PM CDT

gluent.com/event/gluent-webinar-apache-impala-internals-with-tanel-poder

Building an Analytics Platform with Oracle & Hadoop

Speakers: Gerry Moore & Suresh Irukulapati, Vistra Energy

Wednesday, July 26 @ 9 AM CDT

gluent.com/event/gluent-webinar-building-an-integrated-analytics-platform-with-oracle-and-hadoop